Derivation of the EM algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models
DRAFT

Elizabeth Eli Holmes
Northwest Fisheries Science Center, NOAA Fisheries
2725 Montlake Blvd E., Seattle, WA 98112
eli.holmes@noaa.gov
http://faculty.washington.edu/eeholmes

January 30, 2012

Contents

1 Overview 2

2 The EM algorithm 6

3 The unconstrained update equations 10

4 The constrained update equations 24

5 Computing the expectations in the update equations 38

6 Degenerate variance modifications 45

7 Implementation comments 58

8 MARSS R package 59

citation: Holmes, E. E. 2010. Derivation of the EM algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models. Unpublished report. Northwest Fisheries Science Center, NOAA Fisheries, Seattle, WA, USA.

1 Overview

EM algorithms extend likelihood estimation to cases with hidden states, such as when observations are corrupted and the true population size is unobserved. EM algorithms are widely used in engineering and computer science applications. The reader is referred to McLachlan and Krishnan (2008) for general background on EM algorithms and to Harvey (1989) for a discussion of EM algorithms for time-series data. Borman (2009) has a nice tutorial on the EM algorithm. Coding an EM algorithm is not as involved as the following this report might suggest. In most texts, the majority of the steps shown in this technical report would be subsumed under the line "the equations follow directly from the likelihood...". This technical report lays out in detail all of the steps between the likelihood and the EM update equations.

I show first the derivation of the EM algorithm for the unconstrained[1] MARSS model. This EM algorithm was derived by Shumway and Stoffer (1982), but my derivation is in some ways more similar to Ghahramani et al's (Ghahramani and Hinton, 1996; Roweis and Ghahramani, 1999) slightly different presentation. One difference in my presentation and these previous presentations is that I treat the data as a random variable throughout; this means that there are no "special" update equations for the missing values case. I then extend the derivation to the case of a constrained MARSS model where there are fixed and shared elements in the parameter matrices and to the case of a degenerate MARSS model where some processes in the model are deterministic rather than stochastic. An example of a shared value would be a shared drift term (u) across all the random walk processes in a MARSS model. See also Wu et al. (1996) and Zuur et al. (2003) for other examples of the EM algorithm for different classes of constrained MARSS models.

One issue that I do not cover is "identifiability", i.e. does a unique solution exist. For a given MARSS model, you will need to fix some of the parameter elements in order to produce a model with one solution. How to do that depends on how you are using the MARSS model and what specific model you are using. If you are lucky, someone in your field is using a similar type of MARSS model and has already worked out how to constrain the model to ensure identifiability.

Whenever one is working with MARSS models, one should be cognizant that misspecification of the prior on the initial hidden states ($\boldsymbol{x}_0$ or $\boldsymbol{x}_1$) can have catastrophic and difficult to detect effects on your MLE estimates in MARSS models. There is often no sign that something is amiss, except that something seems odd about your parameter estimates. There has been much work on how to avoid these initial conditions effects (see especially literature on VAR state-space models in the economics literature). In our experience, the trouble occurs when the prior on the initial states is inconsistent with the distribution of the initial states that is implied by the MLE model. This often happens when the model implies a specific covariance structure on the initial states. But since you do not know the MLE parameters, you do not know this covariance structure. Using a diffuse prior does not help since your diffuse prior still has some covariance structure (often independence is being imposed). As mentioned above, often

[1] "unconstrained" means that each element in the parameter matrix is estimated and no elements are fixed or shared.

it is very difficult to detect that there is a problem. There are MLE estimates; it is just that these estimates are influenced in a bad way by your prior. One way to detect it is to compare estimates from the EM algorithm versus a Newton-method. If the estimates are quite different, this suggests a prior specification problem because sometimes one or the other algorithm is able/unable to find the MLE when the prior is inconsistent. In some ways the EM algorithm is less sensitive to the prior because it uses the smoothed states in the maximization step. The smoothed states are conditioned on all the data. However, if the prior is inconsistent with the model, the EM algorithm will not (cannot) find the MLE. It is very possible however that it will find parameter estimates that are closer to what you intend (estimates uninfluenced by the prior), but they will not be MLEs. The final section of this report discusses some practical ways to detect the prior problems and to correct or circumvent them.

1.1 The MARSS model

The linear MARSS model with a stochastic initial state[2] is

$$\boldsymbol{x}_t = \mathbf{B}\boldsymbol{x}_{t-1} + \mathbf{u} + \mathbf{w}_t, \text{ where } \mathbf{w}_t \sim \text{MVN}(0, \mathbf{Q}) \tag{1a}$$

$$\boldsymbol{y}_t = \mathbf{Z}\boldsymbol{x}_t + \mathbf{a} + \mathbf{v}_t, \text{ where } \mathbf{v}_t \sim \text{MVN}(0, \mathbf{R}) \tag{1b}$$

$$\boldsymbol{x}_0 \sim \text{MVN}(\xi, \Lambda) \tag{1c}$$

The $\boldsymbol{y}$ equation is called the observation process, and $\boldsymbol{y}_t$ is a $n \times 1$ vector. The $\boldsymbol{x}$ equation is called the state or process equation, and $\boldsymbol{x}_t$ is a $m \times 1$ vector. The equation for $\boldsymbol{x}$ describes a multivariate autoregressive process (also called a random walk or Markov process). The initial state can either defined at $t = 0$, as is done in equation 1, or at $t = 1$. When presenting the MARSS model, I use $t = 0$ but the derivations will show the EM algorithm for both cases. $\mathbf{Q}$ and $\mathbf{R}$ are variance-covariance matrices that specify the stochasticity in the observation and state equations.

This report describes the derivation of an EM algorithm to solve MARSS models, where linear constraints of the form $\beta_i + \beta_{a,i}a + \beta_{b,i}b + \dots$ are placed on the elements in the MARSS parameter matrices. This covers the majority of MARSS models used

[2] 'Stochastic' means the initial state has a distribution rather than a fixed value. Because the process must start somewhere, one needs to specify the initial state as either a distribution or as a parameter. In equation 1, I show the initial state specified as a distribution. However, the derivation will also discuss the case where the initial state is specified as an unknown fixed parameter.

in the literature. Here is an example of a MARSS model with linear constraints:

$$\begin{bmatrix}x_1\\x_2\end{bmatrix}_t = \begin{bmatrix}a&0\\0&2a\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}_{t-1} + \begin{bmatrix}w_1\\w_2\end{bmatrix}_t, \quad \begin{bmatrix}w_1\\w_2\end{bmatrix}_t \sim MVN\left(\begin{bmatrix}0.1\\u+0.1\end{bmatrix}, \begin{bmatrix}q_{11}&q_{12}\\q_{21}&q_{22}\end{bmatrix}\right)$$

$$\begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix}_t = \begin{bmatrix}c & 3c+2d+1\\ c & d\\ c+e+2 & e\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}_t + \begin{bmatrix}v_1\\v_2\\v_3\end{bmatrix}_t,$$
$$\begin{bmatrix}v_1\\v_2\\v_3\end{bmatrix}_t \sim MVN\left(\begin{bmatrix}a_1\\a_2\\0\end{bmatrix}, \begin{bmatrix}r&0&0\\0&2r&0\\0&0&4r\end{bmatrix}\right)$$

$$\begin{bmatrix}x_1\\x_2\end{bmatrix}_0 \sim MVN\left(\begin{bmatrix}\pi\\\pi\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$$

Linear constraints mean that elements of a matrix may be fixed to a specific numerical value or specified as a linear combination of values (which can be shared within a matrix but not shared between matrices).

In the MARSS model, $\boldsymbol{x}$ and $\boldsymbol{y}$ equations describe two stochastic processes. By tradition, one conditions on observations of $\boldsymbol{y}$, and $\boldsymbol{x}$ is treated as completely hidden, hence the name 'hidden Markov process' of which a MARSS model is a special type. However, you could condition on (partial) observations of $\boldsymbol{x}$ and treat $\boldsymbol{y}$ as a (partially) hidden process—with as usual proper constraints to ensure identifiability. Nonetheless in this report, I follow tradition and treat $\boldsymbol{x}$ as hidden and $\boldsymbol{y}$ as (partially) observed. If $\boldsymbol{x}$ is partially observed then the update equations stay the same but the expectations shown in section 5 would be computed conditioned on the partially observed $\boldsymbol{x}$.

1.2 The joint log-likelihood function

Denote the set of all y's and x's from $t = 1$ to T by $\boldsymbol{y}$ and $\boldsymbol{x}$. The joint log-likelihood[3] of $\boldsymbol{y}$ and $\boldsymbol{x}$ can then be written then as follows, where $\boldsymbol{X}_t$ denotes the random variable and $\boldsymbol{x}_t$ is a realization from that random variable (and similarly for $\boldsymbol{Y}_t$):[4]

$$f(\boldsymbol{y},\boldsymbol{x}) = f(\boldsymbol{y}|\boldsymbol{X}=\boldsymbol{x})f(\boldsymbol{x}), \tag{2}$$

where

$$\begin{aligned} f(\boldsymbol{x}) &= f(\boldsymbol{x}_0)\prod_{t=1}^{T} f(\boldsymbol{x}_t|\boldsymbol{X}_1^{t-1}=\boldsymbol{x}_1^{t-1})\\ f(\boldsymbol{y}|\boldsymbol{X}=\boldsymbol{x}) &= \prod_{t=1}^{T} f(\boldsymbol{y}_t|\boldsymbol{X}=\boldsymbol{x}) \end{aligned} \tag{3}$$

[3] This is not the log likelihood output by the Kalman filter. The log likelihood output by the Kalman filter is the $\log \mathbf{L}(\boldsymbol{y};\Theta)$ (notice $\boldsymbol{x}$ does not appear), which is known as the marginal log likelihood.

[4] To alleviate clutter, I have left off subscripts on the f's. To emphasize that the f's represent different density functions, one would often use a subscript showing what parameters are in the functions, i.e. $f(\boldsymbol{x}_t|\boldsymbol{X}_{t-1}=\boldsymbol{x}_{t-1})$ becomes $f_{B,u,Q}(\boldsymbol{x}_t|\boldsymbol{X}_{t-1}=\boldsymbol{x}_{t-1})$.

Thus,

$$\begin{aligned} f(\boldsymbol{y},\boldsymbol{x}) &= \prod_{t=1}^{T} f(\boldsymbol{y}_t|\boldsymbol{X}=\boldsymbol{x}) \times f(\boldsymbol{x}_0) \prod_{t=1}^{T} f(\boldsymbol{x}_t|\boldsymbol{X}_1^{t-1}=\boldsymbol{x}_1^{t-1}) \\ &= \prod_{t=1}^{T} f(\boldsymbol{y}_t|\boldsymbol{X}_t=\boldsymbol{x}_t) \times f(\boldsymbol{x}_0) \prod_{t=1}^{T} f(\boldsymbol{x}_t|\boldsymbol{X}_{t-1}=\boldsymbol{x}_{t-1}). \end{aligned} \tag{4}$$

Here $\boldsymbol{x}_{t1}^{t2}$ denotes the set of $\boldsymbol{x}_t$ from $t = t1$ to $t = t2$ (and thus $\boldsymbol{x}$ is shorthand for $\boldsymbol{x}_1^T$). The third line follows because conditioned on $\boldsymbol{x}$, the $\boldsymbol{y}_t$'s are independent of each other (because the $\mathbf{v}_t$ are independent of each other). In the last line, $\boldsymbol{x}_1^{t-1}$ becomes $\boldsymbol{x}_{t-1}$ from the Markov property of the equation for $\boldsymbol{x}_t$ (equation 1a), and $\boldsymbol{x}$ becomes $\boldsymbol{x}_t$ because $\boldsymbol{y}_t$ depends only on $\boldsymbol{x}_t$ (equation 1b).

Since $(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}=\boldsymbol{x}_{t-1})$ is multivariate normal and $(\boldsymbol{Y}_t|\boldsymbol{X}_t=\boldsymbol{x}_t)$ is multivariate normal (equation 1), we can write down the joint log-likelihood function using the likelihood function for a multivariate normal distribution (Johnson and Wichern, 2007, sec. 4.3).

$$\begin{aligned} \log \mathbf{L}(\boldsymbol{y},\boldsymbol{x};\Theta) = &-\sum_1^T \frac{1}{2}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a})^\top \mathbf{R}^{-1}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a}) - \sum_1^T \frac{1}{2}\log|\mathbf{R}| \\ &-\sum_1^T \frac{1}{2}(\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u})^\top \mathbf{Q}^{-1}(\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u}) - \sum_1^T \frac{1}{2}\log|\mathbf{Q}| \\ &-\frac{1}{2}(\boldsymbol{x}_0 - \xi)^\top \Lambda^{-1}(\boldsymbol{x}_0 - \xi) - \frac{1}{2}\log|\Lambda| - \frac{n}{2}\log 2\pi \end{aligned} \tag{5}$$

n is the number of data points. This is the same as equation 6.64 in Shumway and Stoffer (2006). The above equation is for the case where $\boldsymbol{x}_0$ is stochastic (has a known distribution). However, if we instead treat $\boldsymbol{x}_0$ as fixed but unknown (section 3.4.4 in Harvey, 1989), it is then a parameter and there is no Λ. The likelihood then is slightly different:

$$\begin{aligned} \log \mathbf{L}(\boldsymbol{y},\boldsymbol{x};\Theta) = &-\sum_1^T \frac{1}{2}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a})^\top \mathbf{R}^{-1}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a}) - \sum_1^T \frac{1}{2}\log|\mathbf{R}| \\ &-\sum_1^T \frac{1}{2}(\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u})^\top \mathbf{Q}^{-1}(\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u}) - \sum_1^T \frac{1}{2}\log|\mathbf{Q}| \\ \boldsymbol{x}_0 \equiv \xi & \end{aligned} \tag{6}$$

Note that in this case, $\boldsymbol{x}_0$ is no longer a realization of a random variable $\boldsymbol{X}_0$; it is a fixed (but unknown) parameter. Equation 6 is written as if all the Λ elements are 0 in order to remove clutter, however the MARSS package does not require that all Λ are 0. You can fix some x_0 in $\boldsymbol{x}_0$ and let others have a prior, but you need to make sure the model actually makes sense.

If $\mathbf{R}$ is constant through time, then $\sum_1^T \frac{1}{2}\log|\mathbf{R}|$ in the likelihood equation reduces to $\frac{T}{2}\log|\mathbf{R}|$, however sometimes one needs to includes time-dependent weighting on

$\mathbf{R}$[5]. The same applies to $\sum_1^T \frac{1}{2} \log |\mathbf{Q}|$.

All bolded elements are column vectors (lower case) and matrices (upper case). $\mathbf{A}^\top$ is the transpose of matrix $\mathbf{A}$, $\mathbf{A}^{-1}$ is the inverse of $\mathbf{A}$, and $|\mathbf{A}|$ is the determinant of $\mathbf{A}$. Parameters are non-italic while elements that are slanted are realizations of a random variable ($\boldsymbol{x}$ and $\boldsymbol{y}$ are slated)[6]

1.3 Missing values

In Shumway and Stoffer and other presentations of the EM algorithm for MARSS models (Shumway and Stoffer, 2006; Zuur et al., 2003), the missing values case is treated separately from the non-missing values case. In these derivations, a series of modifications are given for the EM update equations when there are missing values. In my derivation, I present the missing values treatment differently, and there is only one set of update equations and these equations apply in both the missing values and non-missing values cases. My derivation does this by keeping $\mathrm{E}[\boldsymbol{Y}_t|\text{data}]$ and $\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\text{data}]$ in the update equations (much like $\mathrm{E}[\boldsymbol{X}_t|\text{data}]$ is kept in the equations) while Shumway and Stoffer replace these expectations involving $\boldsymbol{Y}_t$ by their values, which depend on whether or not the data are a complete observation of $\boldsymbol{Y}_t$ with no missing values. Section 5 shows how to compute the expectations involving $\boldsymbol{Y}_t$ when the data are an incomplete observation of $\boldsymbol{Y}_t$.

2 The EM algorithm

The EM algorithm cycles iteratively between an expectation step (the integration in the equation) followed by a maximization step (the arg max in the equation):

$$\Theta_{j+1} = \arg\max_{\Theta} \int_{\boldsymbol{x}} \int_{\boldsymbol{y}} \log \mathbf{L}(\boldsymbol{x},\boldsymbol{y};\Theta) f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j) d\boldsymbol{x} d\boldsymbol{y} \tag{7}$$

$\boldsymbol{Y}(1)$ indicates those $\boldsymbol{Y}$ that have an observation and $\boldsymbol{y}(1)$ are the actual observations. Note that Θ and Θ_j are different. If Θ consists of multiple parameters, we can also break this down into smaller steps. Let $\Theta = \{\alpha, \beta\}$, then

$$\alpha_{j+1} = \arg\max_{\alpha} \int_{\boldsymbol{x}} \int_{\boldsymbol{y}} \log \mathbf{L}(\boldsymbol{x},\boldsymbol{y},\beta_j;\alpha) f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \alpha_j, \beta_j) d\boldsymbol{x} d\boldsymbol{y} \tag{8}$$

Now the maximization is only over α, the part that appears after the ";" in the log-likelihood.

Expectation step The integral that appears in equation (7) is an expectation. The first step in the EM algorithm is to compute this expectation. This will involve computing expectations like $\mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1), \Theta_j]$ and $\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1), \Theta_j]$. The j subscript on Θ denotes that these are the parameters at iteration j of the algorithm.

[5] If for example, one wanted to include a temporally dependent weighting on $\mathbf{R}$ replace $|\mathbf{R}|$ with $|\alpha_t \mathbf{R}| = \alpha_t^n |\mathbf{R}|$, where α_t is the weighting at time t and is fixed not estimated.

[6] In matrix algebra, a capitol bolded letter indicates a matrix. Unfortunately in statistics, the capitol letter convention is used for random variables. Fortunately, this derivation does not need to reference random variables except indirectly when using expectations. Thus, I use capitols to refer to matrices not random variables. The one exception is the reference to $\boldsymbol{X}$ and in this case a bolded *slanted* capitol is used.

Maximization step: A new parameter set Θ_{j+1} is computed by finding the parameters that maximize the *expected* log-likelihood function (the part in the integral) with respect to Θ. The equations that give the parameters for the next iteration $(j+1)$ are called the update equations and this report is devoted to the derivation of these update equations.

After one iteration of the expectation and maximization steps, the cycle is then repeated. New expectations are computed using Θ_{j+1}, and then a new set of parameters Θ_{j+2} is generated. This cycle is continued until the likelihood no longer increases more than a specified tolerance level. This algorithm is guaranteed to increase in likelihood at each iteration (if it does not, it means there is an error in one's update equations). The algorithm must be started from an initial set of parameter values Θ_1. The algorithm is not particularly sensitive to the initial conditions but the surface could definitely be multi-modal and have local maxima. See section 7 on using Monte Carlo initialization to ensure that the global maximum is found.

2.1 The expected log-likelihood function

The function that is maximized in the "M" step is the expected value of the log-likelihood function. This expectation is conditioned on two things: 1) the observed $\boldsymbol{Y}$'s which are denoted $\boldsymbol{Y}(1)$ and which are equal to the fixed values $\boldsymbol{y}(1)$ and 2) the parameter set Θ_j. Note that since there may be missing values in the data, $\boldsymbol{Y}(1)$ can be a subset of $\boldsymbol{Y}$, that is, only some $\boldsymbol{Y}$ have a corresponding $\boldsymbol{y}$ value at time t. Mathematically what we are doing is $\text{E}_{\mathbf{XY}}[g(\boldsymbol{X},\boldsymbol{Y})|\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j]$. This is a multivariate conditional expectation because $\boldsymbol{X},\boldsymbol{Y}$ is multivariate (a $m \times n \times T$ vector). The function $g(\Theta)$ that we are taking the expectation of is $\log \mathbf{L}(\boldsymbol{Y},\boldsymbol{X};\Theta)$. Note that $g(\Theta)$ is a random variable involving the random variables, $\boldsymbol{X}$ and $\boldsymbol{Y}$, while $\log \mathbf{L}(\boldsymbol{y},\boldsymbol{x};\Theta)$ is not a random variable but rather a specific value since $\boldsymbol{y}$ and $\boldsymbol{x}$ are a set of specific values.

We denote this expected log-likelihood by Ψ. Using the log likelihood equation (5) and expanding out all the terms, we can write out Ψ as:

$$
\begin{aligned}
&\mathrm{E}_{\mathbf{XY}}[\log \mathbf{L}(\boldsymbol{Y},\boldsymbol{X};\Theta);\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j]=\Psi=\\
&-\frac{1}{2}\sum_1^T\Bigg(\mathrm{E}[\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\boldsymbol{Y}_t]-\mathrm{E}[\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t]-\mathrm{E}[(\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\boldsymbol{Y}_t]\\
&-\mathrm{E}[\mathbf{a}^\top\mathbf{R}^{-1}\boldsymbol{Y}_t]-\mathrm{E}[\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\mathbf{a}]+\mathrm{E}[(\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t]\\
&+\mathrm{E}[\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t]+\mathrm{E}[(\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\mathbf{a}]+\mathrm{E}[\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{a}]\Bigg)-\frac{T}{2}\log|\mathbf{R}|\\
&-\frac{1}{2}\sum_1^T\Bigg(\mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\boldsymbol{X}_t]-\mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}]\\
&-\mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\boldsymbol{X}_t]-\mathrm{E}[\mathbf{u}^\top\mathbf{Q}^{-1}\boldsymbol{X}_t]-\mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{u}]\\
&+\mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}]+\mathrm{E}[\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}]\\
&+\mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{u}]+\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{u}\Bigg)-\frac{T}{2}\log|\mathbf{Q}|\\
&-\frac{1}{2}\Bigg(\mathrm{E}[\boldsymbol{X}_0^\top\mathbf{V}_0^{-1}\boldsymbol{X}_0]-\mathrm{E}[\xi^\top\Lambda^{-1}\boldsymbol{X}_0]\\
&-\mathrm{E}[\boldsymbol{X}_0^\top\Lambda^{-1}\xi]+\xi^\top\Lambda^{-1}\xi\Bigg)-\frac{1}{2}\log|\Lambda|-\frac{n}{2}\log\pi
\end{aligned}
\tag{9}
$$

All the E[ ] appearing here denote $\mathrm{E}_{\mathbf{XY}}[g()|\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j]$. In the rest of the derivation, I drop the conditional and the XY subscript on E to remove clutter, but it is important to remember that whenever E appears, it refers to a specific conditional multivariate expectation. If $\boldsymbol{x}_0$ is treated as fixed, then $\boldsymbol{X}_0=\xi$ and the last two lines involving Λ are dropped.

Keep in mind that Θ and Θ_j are different. Θ is a parameter appearing in function $g(\boldsymbol{X},\boldsymbol{Y},\Theta)$. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random variables which means that $g(\boldsymbol{X},\boldsymbol{Y},\Theta)$ is a random variable. We take the expectation of $g(\boldsymbol{X},\boldsymbol{Y},\Theta)$, meaning we take integral over the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$. We need to specify what that distribution is and the conditioning on Θ_j is specifying that. This conditioning affects the value of the expectation of $g(\boldsymbol{X},\boldsymbol{Y},\Theta)$, but it does not affect the value of Θ, which are the $\mathbf{R}$, $\mathbf{Q}$, $\mathbf{u}$, etc. values on the right side. We will first take the expectation of $g(\boldsymbol{X},\boldsymbol{Y},\Theta)$ conditioned on Θ_j (using integration) and then take the differential of that expectation with respect to Θ.

I will reference the expected log-likelihood throughout the derivation of the update equations. It could be written more concisely, but for deriving the update equations, I will keep it in this verbose form. The goal is to find the Θ that maximizes this expectation and this becomes the new parameter set for the $j+1$ iteration of the EM algorithm. The equations to compute these new parameters are termed the update equations.

Table 1: Notes on multivariate expectations. For the following examples, let $\boldsymbol{X}$ be a vector of length three, X_1, X_2, X_3. $f()$ is the probability distribution function (pdf). C is a constant (not a random variable).

$$
\begin{array}{l}
\mathrm{E}_X[g(\boldsymbol{X})] = \int\int\int g(\boldsymbol{x}) f(x_1,x_2,x_3)dx_1dx_2dx_3 \\
\mathrm{E}_X[X_1] = \int\int\int x_1 f(x_1,x_2,x_3)dx_1dx_2dx_3 = \int x_1 f(x_1)dx_1 = \mathrm{E}[X_1] \\
\mathrm{E}_X[X_1+X_2] = \mathrm{E}_X[X_1] + \mathrm{E}_X[X_2] \\
\mathrm{E}_X[X_1+C] = \mathrm{E}_X[X_1] + C \\
\mathrm{E}_X[CX_1] = C\,\mathrm{E}_X[X_1] \\
\mathrm{E}_X[X_1|X_1 = x_1] = x_1 \\
\mathrm{E}_X[\boldsymbol{X}|\boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{x}
\end{array}
$$

2.2 The expectations used in the derivation

The following expectations appear frequently in the update equations and are given special names[7]:

$$
\begin{aligned}
\widetilde{\mathbf{x}}_t &= \mathrm{E}_{\mathbf{XY}}[\boldsymbol{X}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j] && \text{(10a)} \\
\widetilde{\mathbf{y}}_t &= \mathrm{E}_{\mathbf{XY}}[\boldsymbol{Y}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j] && \text{(10b)} \\
\widetilde{\mathbf{P}}_t &= \mathrm{E}_{\mathbf{XY}}[\boldsymbol{X}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j] && \text{(10c)} \\
\widetilde{\mathbf{P}}_{t,t-1} &= \mathrm{E}_{\mathbf{XY}}[\boldsymbol{X}_t\boldsymbol{X}_{t-1}^\top|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j] && \text{(10d)} \\
\widetilde{\mathbf{V}}_t &= \mathrm{var}_{XY}[\boldsymbol{X}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j] = \widetilde{\mathbf{P}}_t - \widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^\top && \text{(10e)} \\
\widetilde{\mathbf{O}}_t &= \mathrm{E}_{\mathbf{XY}}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j] && \text{(10f)} \\
\widetilde{\mathbf{W}}_t &= \mathrm{var}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j] = \widetilde{\mathbf{O}}_t - \widetilde{\mathbf{y}}_t\widetilde{\mathbf{y}}_t^\top && \text{(10g)} \\
\widetilde{\mathbf{yx}}_t &= \mathrm{E}_{\mathbf{XY}}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j] && \text{(10h)} \\
\widetilde{\mathbf{yx}}_{t,t-1} &= \mathrm{E}_{\mathbf{XY}}[\boldsymbol{Y}_t\boldsymbol{X}_{t-1}^\top|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j] && \text{(10i)}
\end{aligned}
$$

The subscript on the expectation, E, denotes that this is a multivariate expectation taken over $\boldsymbol{X}$ and $\boldsymbol{Y}$. The right sides of equations (10e) and (10g) arise from the computational formula for variance and covariance:

$$
\begin{aligned}
\mathrm{var}[X] &= \mathrm{E}[XX^\top] - \mathrm{E}[X]\,\mathrm{E}[X]^\top && \text{(11)} \\
\mathrm{cov}[X,Y] &= \mathrm{E}[XY^\top] - \mathrm{E}[X]\,\mathrm{E}[Y]^\top. && \text{(12)}
\end{aligned}
$$

Section 5 shows how to compute the expectations in equation 10.

[7] This notation is different than what you see in Shumway and Stoffer (2006), section 6.2. What I call $\widetilde{\mathbf{V}}_t$, they refer to as P_t^n, and my $\widetilde{\mathbf{P}}_t$ would be $P_t^n + \widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t'$ in their notation.

3 The unconstrained update equations

In this section, I show the derivation of the update equations when all elements of a parameter matrix are estimated and are all allowed to be different; these are similar to the update equations one will see in Shumway and Stoffer's text. Section 4 shows the update equations when there are fixed or shared values in the parameter matrices, i.e. the constrained update equations.

To derive the update equations, one must find the Θ, where Θ is comprised of the MARSS parameters $\mathbf{B}$, $\mathbf{u}$, $\mathbf{Q}$, $\mathbf{Z}$, $\mathbf{a}$, $\mathbf{R}$, ξ, and Λ, that maximizes Ψ (equation 9) by partial differentiation of Ψ with respect to Θ. However, I will be using the EM equation where one maximizes each parameter matrix in Θ one-by-one (equation 8). In this case, the parameters that are not being maximized are set at their iteration j values, and then one takes the derivative of Ψ with respect to the parameter of interest. Then solve for the parameter value that sets the partial derivative to zero. The partial differentiation is with respect to each individual parameter element, for example each u_k in the vector $\mathbf{u}$. The idea is to single out those terms in equation (9) that involve u_k (say), differentiate by u_k, set this to zero and solve for u_k. This gives the new u_k that maximizes the partial derivative with respect to u_k of the expected log-likelihood. Matrix calculus gives us a way to jointly maximize Ψ with respect to all elements (not just element k) in a parameter vector or matrix.

3.1 Matrix calculus need for the derivation

Before commencing, some definitions from matrix calculus will be needed. The partial derivative of a scalar (Ψ is a scalar) with respect to some column vector $\mathbf{b}$ (which has elements $b_1, b_2 \dots$) is

$$\frac{\partial\Psi}{\partial\mathbf{b}} = \begin{bmatrix} \frac{\partial\Psi}{\partial b_1} & \frac{\partial\Psi}{\partial b_2} & \dots & \frac{\partial\Psi}{\partial b_n} \end{bmatrix}$$

Note that the derivative of a column vector $\mathbf{b}$ is a row vector. The partial derivatives of a scalar with respect to some $n \times n$ matrix $\mathbf{B}$ is

$$\frac{\partial\Psi}{\partial\mathbf{B}} = \begin{bmatrix} \frac{\partial\Psi}{\partial b_{1,1}} & \frac{\partial\Psi}{\partial b_{2,1}} & \dots & \frac{\partial\Psi}{\partial b_{n,1}} \\ \frac{\partial\Psi}{\partial b_{1,2}} & \frac{\partial\Psi}{\partial b_{2,2}} & \dots & \frac{\partial\Psi}{\partial b_{n,2}} \\ \dots & \dots & \dots & \dots \\ \frac{\partial\Psi}{\partial b_{1,n}} & \frac{\partial\Psi}{\partial b_{2,n}} & \dots & \frac{\partial\Psi}{\partial b_{n,n}} \end{bmatrix}$$

Note that the indexing is interchanged; $\partial\Psi/\partial b_{i,j} = \big[\partial\Psi/\partial\mathbf{B}\big]_{j,i}$. For $\mathbf{Q}$ and $\mathbf{R}$, this is unimportant because they are variance-covariance matrices and are symmetric. For $\mathbf{B}$ and $\mathbf{Z}$, one must be careful because these may not be symmetric.

Table 2: Derivatives of a scalar with respect to vectors and matrices. In the following $\mathbf{a}$ and $\mathbf{c}$ are $n \times 1$ column vectors, $\mathbf{b}$ and $\mathbf{d}$ are $m \times 1$ column vectors, $\mathbf{D}$ is a $n \times m$ matrix, $\mathbf{C}$ is a $n \times n$ matrix, and $\mathbf{A}$ is a diagonal $n \times n$ matrix (0s on the off-diagonals). $\mathbf{C}^{-1}$ is the inverse of $\mathbf{C}$, $\mathbf{C}^\top$ is the transpose of $\mathbf{C}$, $\mathbf{C}^{-\top} = \big(\mathbf{C}^{-1}\big)^\top = \big(\mathbf{C}^\top\big)^{-1}$, and $|\mathbf{C}|$ is the determinant of $\mathbf{C}$. Note, all the numerators in the differentials reduce to scalars.

$$\partial(\mathbf{a}^\top\mathbf{c})/\partial\mathbf{a} = \partial(\mathbf{c}^\top\mathbf{a})/\partial\mathbf{a} = \mathbf{c}^\top \tag{13}$$

$$\begin{aligned}
&\partial(\mathbf{a}^\top\mathbf{D}\mathbf{b})/\partial\mathbf{D} = \partial(\mathbf{b}^\top\mathbf{D}^\top\mathbf{a})/\partial\mathbf{D} = \mathbf{b}\mathbf{a}^\top\\
&\partial(\mathbf{a}^\top\mathbf{D}\mathbf{b})/\partial\,\mathrm{vec}(\mathbf{D}) = \partial(\mathbf{b}^\top\mathbf{D}^\top\mathbf{a})/\partial\,\mathrm{vec}(\mathbf{D}) = \big(\mathrm{vec}(\mathbf{b}\mathbf{a}^\top)\big)^\top
\end{aligned} \tag{14}$$

$$\begin{aligned}
&\partial(\log|\mathbf{C}|)/\partial\mathbf{C} = -\partial(\log|\mathbf{C}^{-1}|)/\partial\mathbf{C} = (\mathbf{C}^\top)^{-1} = \mathbf{C}^{-\top}\\
&\partial(\log|\mathbf{C}|)/\partial\,\mathrm{vec}(\mathbf{C}) = \big(\mathrm{vec}(\mathbf{C}^{-\top})\big)^\top
\end{aligned} \tag{15}$$

$$\begin{aligned}
&\partial(\mathbf{b}^\top\mathbf{D}^\top\mathbf{C}\mathbf{D}\mathbf{d})/\partial\mathbf{D} = \mathbf{d}\mathbf{b}^\top\mathbf{D}^\top\mathbf{C} + \mathbf{b}\mathbf{d}^\top\mathbf{D}^\top\mathbf{C}^\top\\
&\partial(\mathbf{b}^\top\mathbf{D}^\top\mathbf{C}\mathbf{D}\mathbf{d})/\partial\,\mathrm{vec}(\mathbf{D}) = \big(\mathrm{vec}(\mathbf{d}\mathbf{b}^\top\mathbf{D}^\top\mathbf{C} + \mathbf{b}\mathbf{d}^\top\mathbf{D}^\top\mathbf{C}^\top)\big)^\top\\
&\text{If } \mathbf{b} = \mathbf{d} \text{ and } \mathbf{C} \text{ is symmetric then the sum reduces to } 2\mathbf{b}\mathbf{b}^\top\mathbf{D}^\top\mathbf{C}
\end{aligned} \tag{16}$$

$$\partial(\mathbf{a}^\top\mathbf{C}\mathbf{a})/\partial\mathbf{a} = \partial(\mathbf{a}\mathbf{C}^\top\mathbf{a}^\top)/\partial\mathbf{a} = 2\mathbf{a}^\top\mathbf{C} \tag{17}$$

$$\begin{aligned}
&\partial(\mathbf{a}^\top\mathbf{C}^{-1}\mathbf{c})/\partial\mathbf{C} = -\mathbf{C}^{-1}\mathbf{a}\mathbf{c}^\top\mathbf{C}^{-1}\\
&\partial(\mathbf{a}^\top\mathbf{C}^{-1}\mathbf{c})/\partial\,\mathrm{vec}(\mathbf{C}) = -\big(\mathrm{vec}(\mathbf{C}^{-1}\mathbf{a}\mathbf{c}^\top\mathbf{C}^{-1})\big)^\top
\end{aligned} \tag{18}$$

A number of derivatives of a scalar with respect to vectors and matrices will be needed in the derivation and are shown in table 2. In the table, both the vectorized and non-vectorized versions are shown. The vectorized version of a matrix $\mathbf{D}$ with dimension $n \times m$ is

$$\mathrm{vec}(\mathbf{D}_{n,m}) \equiv \begin{bmatrix} d_{1,1}\\ \cdots\\ d_{n,1}\\ d_{1,2}\\ \cdots\\ d_{n,2}\\ \cdots\\ d_{1,m}\\ \cdots\\ d_{n,m} \end{bmatrix}$$

3.2 The update equation for u (unconstrained)

Take the partial derivative of Ψ with respect to $\mathbf{u}$, which is a $m \times 1$ column vector. All parameters other than $\mathbf{u}$ are fixed to constant values (because partial derivation is being done). Since the derivative of a constant is 0, terms not involving $\mathbf{u}$ will equal 0 and drop out. Taking the derivative to equation (9) with respect to $\mathbf{u}$:

$$\begin{aligned}\partial\Psi/\partial\mathbf{u} = -\frac{1}{2}\sum_{t=1}^{T}\bigg(&-\partial(\mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{u}])/\partial\mathbf{u} - \partial(\mathrm{E}[\mathbf{u}^\top\mathbf{Q}^{-1}\boldsymbol{X}_t])/\partial\mathbf{u}\\ &+\partial(\mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{u}])/\partial\mathbf{u} + \partial(\mathrm{E}[\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}])/\partial\mathbf{u}\\ &+\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{u}\bigg)\end{aligned} \tag{19}$$

The parameters can be moved out of the expectations and then the relations (13) and (17) are used to take the derivative.

$$\begin{aligned}\partial\Psi/\partial\mathbf{u} = -\frac{1}{2}\sum_{t=1}^{T}\bigg(&-\mathrm{E}[\boldsymbol{X}_t]^\top\mathbf{Q}^{-1} - (\mathbf{Q}^{-1}\mathrm{E}[\boldsymbol{X}_t])^\top\\ &+(\mathbf{B}^\top\mathrm{E}[\boldsymbol{X}_{t-1}])^\top\mathbf{Q}^{-1} + (\mathbf{Q}^{-1}\mathbf{B}\mathrm{E}[\boldsymbol{X}_{t-1}])^\top + 2\mathbf{u}^\top\mathbf{Q}^{-1}\bigg)\end{aligned} \tag{20}$$

This also uses $\mathbf{Q}^{-1} = (\mathbf{Q}^{-1})^\top$. This can then be reduced to

$$\partial\Psi/\partial\mathbf{u} = \sum_{t=1}^{T}\big(\mathrm{E}[\boldsymbol{X}_t]^\top\mathbf{Q}^{-1} - \mathrm{E}[\boldsymbol{X}_{t-1}]^\top\mathbf{B}^\top\mathbf{Q}^{-1} - \mathbf{u}^\top\mathbf{Q}^{-1}\big) \tag{21}$$

Set the left side to zero (a $1 \times m$ matrix of zeros) and transpose the whole equation. $\mathbf{Q}^{-1}$ cancels out[8] by multiplying on the left by $\mathbf{Q}$ (left since the whole equation was just transposed), giving

$$\mathbf{0} = \sum_{t=1}^{T}\big(\mathrm{E}[\boldsymbol{X}_t] - \mathbf{B}\mathrm{E}[\boldsymbol{X}_{t-1}] - \mathbf{u}\big) = \sum_{t=1}^{T}\big(\mathrm{E}[\boldsymbol{X}_t] - \mathbf{B}\mathrm{E}[\boldsymbol{X}_{t-1}]\big) - T\mathbf{u} \tag{22}$$

Solving for $\mathbf{u}$ and replacing the expectations with their names from equation 10, gives us the new $\mathbf{u}$ that maximizes Ψ,

$$\mathbf{u}_{j+1} = \frac{1}{T}\sum_{t=1}^{T}\big(\widetilde{\mathbf{x}}_t - \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\big) \tag{23}$$

3.3 The update equation for B (unconstrained)

Take the derivative of Ψ with respect to $\mathbf{B}$. Terms not involving $\mathbf{B}$, equal 0 and drop out. I have put the E outside the partials by noting that $\partial(\mathrm{E}[h(\boldsymbol{X}_t,\mathbf{B})])/\partial\mathbf{B} =$

[8] $\mathbf{Q}$ is a variance-covariance matrix and is invertible. $\mathbf{Q}^{-1}\mathbf{Q} = \mathbf{I}$, the identity matrix.

$\mathrm{E}[\partial(h(\boldsymbol{X}_t,\mathbf{B}))/\partial\mathbf{B}]$ since the expectation is conditioned on $\mathbf{B}_j$ not $\mathbf{B}$.

$$
\begin{aligned}
\partial\Psi/\partial\mathbf{B} &= -\frac{1}{2}\sum_{t=1}^{T}\Big(-\mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1})/\partial\mathbf{B}] \\
&\quad -\mathrm{E}[\partial((\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\boldsymbol{X}_t)/\partial\mathbf{B}] + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}(\mathbf{B}\boldsymbol{X}_{t-1}))/\partial\mathbf{B}] \\
&\quad + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{B}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1})/\partial\mathbf{B}]\Big) \\
&= -\frac{1}{2}\sum_{t=1}^{T}\Big(-\mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}])/\partial\mathbf{B}] \\
&\quad -\mathrm{E}[\partial(\boldsymbol{X}_{t-1}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\boldsymbol{X}_t)/\partial\mathbf{B}] + \mathrm{E}[\partial(\boldsymbol{X}_{t-1}^\top\mathbf{B}^\top\mathbf{Q}^{-1}(\mathbf{B}\boldsymbol{X}_{t-1}))/\partial\mathbf{B}] \\
&\quad + \mathrm{E}[\partial(\boldsymbol{X}_{t-1}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{B}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1})/\partial\mathbf{B}\Big)]
\end{aligned}
\tag{24}
$$

After pulling the constants out of the expectations, we use relations (14) and (16) to take the derivative and note that $\mathbf{Q}^{-1} = (\mathbf{Q}^{-1})^\top$:

$$
\begin{aligned}
\partial\Psi/\partial\mathbf{B} &= -\frac{1}{2}\sum_{t=1}^{T}\Big(-\mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_t^\top]\mathbf{Q}^{-1} - \mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_t^\top]\mathbf{Q}^{-1} \\
&\quad + 2\,\mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top]\mathbf{B}^\top\mathbf{Q}^{-1} + \mathrm{E}[\boldsymbol{X}_{t-1}]\mathbf{u}^\top\mathbf{Q}^{-1} + \mathrm{E}[\boldsymbol{X}_{t-1}]\mathbf{u}^\top\mathbf{Q}^{-1}\Big)
\end{aligned}
\tag{25}
$$

This can be reduced to

$$
\begin{aligned}
\partial\Psi/\partial\mathbf{B} &= -\frac{1}{2}\sum_{t=1}^{T}\Big(-2\,\mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_t^\top]\mathbf{Q}^{-1} \\
&\quad + 2\,\mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top]\mathbf{B}^\top\mathbf{Q}^{-1} + 2\,\mathrm{E}[\boldsymbol{X}_{t-1}]\mathbf{u}^\top\mathbf{Q}^{-1}\Big)
\end{aligned}
\tag{26}
$$

Set the left side to zero (an $m \times m$ matrix of zeros), cancel out $\mathbf{Q}^{-1}$ by multiplying by $\mathbf{Q}$ on the right, get rid of the -1/2, and transpose the whole equation to give

$$
\begin{aligned}
\mathbf{0} &= \sum_{t=1}^{T}\big(\mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_{t-1}^\top] - \mathbf{B}\,\mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top] - \mathbf{u}\,\mathrm{E}[\boldsymbol{X}_{t-1}^\top]\big) \\
&= \sum_{t=1}^{T}\big(\widetilde{\mathbf{P}}_{t,t-1} - \mathbf{B}\widetilde{\mathbf{P}}_{t-1} - \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top\big)
\end{aligned}
\tag{27}
$$

The last line replaced the expectations with their names shown in equation (10). Solving for $\mathbf{B}$ and noting that $\widetilde{\mathbf{P}}_{t-1}$ is like a variance-covariance matrix and is invertible, gives us the new $\mathbf{B}$ that maximizes Ψ,

$$
\mathbf{B}_{j+1} = \Big(\sum_{t=1}^{T}\big(\widetilde{\mathbf{P}}_{t,t-1} - \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top\big)\Big)\Big(\sum_{t=1}^{T}\widetilde{\mathbf{P}}_{t-1}\Big)^{-1}
\tag{28}
$$

Because all the equations above also apply to block-diagonal matrices, the derivation immediately generalizes to the case where $\mathbf{B}$ is an unconstrained block diagonal matrix:

$$\mathbf{B} = \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} & 0 & 0 & 0 & 0 & 0 \\ b_{2,1} & b_{2,2} & b_{2,3} & 0 & 0 & 0 & 0 & 0 \\ b_{3,1} & b_{3,2} & b_{3,3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_{4,4} & b_{4,5} & 0 & 0 & 0 \\ 0 & 0 & 0 & b_{5,4} & b_{5,5} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & b_{6,6} & b_{6,7} & b_{6,8} \\ 0 & 0 & 0 & 0 & 0 & b_{7,6} & b_{7,7} & b_{7,8} \\ 0 & 0 & 0 & 0 & 0 & b_{8,6} & b_{8,7} & b_{8,8} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 & 0 & 0 \\ 0 & \mathbf{B}_2 & 0 \\ 0 & 0 & \mathbf{B}_3 \end{bmatrix}$$

For the block diagonal $\mathbf{B}$,

$$\mathbf{B}_{i,j+1} = \bigg(\sum_{t=1}^{T} \big(\widetilde{\mathbf{P}}_{t,t-1} - \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top \big) \bigg)_i \bigg(\sum_{t=1}^{T} \widetilde{\mathbf{P}}_{t-1} \bigg)_i^{-1} \tag{29}$$

where the subscript i means to take the parts of the matrices that are analogous to $\mathbf{B}_i$; take the whole part within the parentheses not the individual matrices inside the parentheses. If $\mathbf{B}_i$ is comprised of rows a to b and columns c to d of matrix $\mathbf{B}$, then take rows a to b and columns c to d of the matrices subscripted by i in equation (29).

3.4 The update equation for Q (unconstrained)

The usual way to do this derivation is to use what is known as the "trace trick" which will pull the $\mathbf{Q}^{-1}$ out to the left of the $\mathbf{c}^\top\mathbf{Q}^{-1}\mathbf{b}$ terms which appear in the likelihood (9). Here I'm showing a less elegant derivation that plods step by step through each of the likelihood terms. Take the derivative of Ψ with respect to $\mathbf{Q}$. Terms not involving $\mathbf{Q}$ equal 0 and drop out. Again the expectations are placed outside the partials by noting that $\partial(\mathrm{E}[h(\boldsymbol{X}_t,\mathbf{Q})])/\partial\mathbf{Q} = \mathrm{E}[\partial(h(\boldsymbol{X}_t,\mathbf{Q}))/\partial\mathbf{Q}]$.

$$\begin{aligned} \partial\Psi/\partial\mathbf{Q} = -\frac{1}{2}\sum_{t=1}^{T} \bigg(& \mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\boldsymbol{X}_t)/\partial\mathbf{Q}] - \mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1})/\partial\mathbf{Q}] \\ & - \mathrm{E}[\partial((\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\boldsymbol{X}_t)/\partial\mathbf{Q}] - \mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{Q}] \\ & - \partial(\mathrm{E}[\mathbf{u}^\top\mathbf{Q}^{-1}\boldsymbol{X}_t)/\partial\mathbf{Q}] + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1})/\partial\mathbf{Q}] \\ & + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{Q}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1})/\partial\mathbf{Q}] \\ & + \partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{Q} \bigg) - \partial\bigg(\frac{T}{2}\log|\mathbf{Q}|\bigg)/\partial\mathbf{Q} \end{aligned} \tag{30}$$

The relations (18) and (15) are used to do the differentiation. Notice that all the terms in the summation are of the form $\mathbf{c}^\top\mathbf{Q}^{-1}\mathbf{b}$, and thus after differentiation, all the $\mathbf{c}^\top\mathbf{b}$ terms can be grouped inside one set of parentheses. Also there is a minus that comes from equation (18) and it cancels out the minus in front of the initial $-1/2$.

$$
\begin{aligned}
\partial\Psi/\partial\mathbf{Q} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{Q}^{-1}\bigg(&\operatorname{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top] - \operatorname{E}[\boldsymbol{X}_t(\mathbf{B}\boldsymbol{X}_{t-1})^\top] - \operatorname{E}[\mathbf{B}\boldsymbol{X}_{t-1}\boldsymbol{X}_t^\top]\\
&- \operatorname{E}[\boldsymbol{X}_t\mathbf{u}^\top] - \operatorname{E}[\mathbf{u}\boldsymbol{X}_t^\top] + \operatorname{E}[\mathbf{B}\boldsymbol{X}_{t-1}(\mathbf{B}\boldsymbol{X}_{t-1})^\top] + \operatorname{E}[\mathbf{B}\boldsymbol{X}_{t-1}\mathbf{u}^\top]\\
&+ \operatorname{E}[\mathbf{u}(\mathbf{B}\boldsymbol{X}_{t-1})^\top] + \mathbf{u}\mathbf{u}^\top\bigg)\mathbf{Q}^{-1} - \frac{T}{2}\mathbf{Q}^{-1}
\end{aligned}
\tag{31}
$$

Pulling the parameters out of the expectations and using $(\mathbf{B}\boldsymbol{X}_t)^\top = \boldsymbol{X}_t^\top\mathbf{B}^\top$, we have

$$
\begin{aligned}
\partial\Psi/\partial\mathbf{Q} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{Q}^{-1}\bigg(&\operatorname{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top] - \operatorname{E}[\boldsymbol{X}_t\boldsymbol{X}_{t-1}^\top]\mathbf{B}^\top - \mathbf{B}\operatorname{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_t^\top]\\
&- \operatorname{E}[\boldsymbol{X}_t]\mathbf{u}^\top - \mathbf{u}\operatorname{E}[\boldsymbol{X}_t^\top] + \mathbf{B}\operatorname{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top]\mathbf{B}^\top + \mathbf{B}\operatorname{E}[\boldsymbol{X}_{t-1}]\mathbf{u}^\top\\
&+ \mathbf{u}\operatorname{E}[\boldsymbol{X}_{t-1}^\top]\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\bigg)\mathbf{Q}^{-1} - \frac{T}{2}\mathbf{Q}^{-1}
\end{aligned}
\tag{32}
$$

The partial derivative is then rewritten in terms of the Kalman smoother output:

$$
\begin{aligned}
\partial\Psi/\partial\mathbf{Q} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{Q}^{-1}\bigg(&\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}\widetilde{\mathbf{x}}_t^\top\\
&+ \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\bigg)\mathbf{Q}^{-1} - \frac{T}{2}\mathbf{Q}^{-1}
\end{aligned}
\tag{33}
$$

Setting this to zero (a $m \times m$ matrix of zeros), $\mathbf{Q}^{-1}$ is canceled out by multiplying by $\mathbf{Q}$ twice, once on the left and once on the right and the $1/2$ is removed.

$$
\begin{aligned}
\mathbf{0} = \sum_{t=1}^{T}\bigg(&\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}\widetilde{\mathbf{x}}_t^\top\\
&+ \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\bigg) - T\mathbf{Q}
\end{aligned}
\tag{34}
$$

We can then solve for $\mathbf{Q}$, giving us the new $\mathbf{Q}$ that maximizes Ψ,

$$
\begin{aligned}
\mathbf{Q}_{j+1} = \frac{1}{T}\sum_{t=1}^{T}\bigg(&\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}\widetilde{\mathbf{x}}_t^\top\\
&+ \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\bigg)
\end{aligned}
\tag{35}
$$

This derivation immediately generalizes to the case where $\mathbf{Q}$ is a block diagonal matrix:

$$\mathbf{Q} = \begin{bmatrix} q_{1,1} & q_{1,2} & q_{1,3} & 0 & 0 & 0 & 0 & 0 \\ q_{1,2} & q_{2,2} & q_{2,3} & 0 & 0 & 0 & 0 & 0 \\ q_{1,3} & q_{2,3} & q_{3,3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & q_{4,4} & q_{4,5} & 0 & 0 & 0 \\ 0 & 0 & 0 & q_{4,5} & q_{5,5} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & q_{6,6} & q_{6,7} & q_{6,8} \\ 0 & 0 & 0 & 0 & 0 & q_{6,7} & q_{7,7} & q_{7,8} \\ 0 & 0 & 0 & 0 & 0 & q_{6,8} & q_{7,8} & q_{8,8} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_1 & 0 & 0 \\ 0 & \mathbf{Q}_2 & 0 \\ 0 & 0 & \mathbf{Q}_3 \end{bmatrix}$$

In this case,

$$\begin{split}\mathbf{Q}_{i,j+1} = \frac{1}{T}\sum_{t=1}^{T}\Big(\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}\widetilde{\mathbf{x}}_t^\top \\ + \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\Big)_i\end{split} \tag{36}$$

where the subscript i means take the elements of the matrix (in the big parentheses) that are analogous to $\mathbf{Q}_i$; take the whole part within the parentheses not the individual matrices inside the parentheses). If $\mathbf{Q}_i$ is comprised of rows a to b and columns c to d of matrix $\mathbf{Q}$, then take rows a to b and columns c to d of matrices subscripted by i in equation (36).

By the way, $\mathbf{Q}$ is never really unconstrained since it is a variance-covariance matrix and the upper and lower triangles are shared. However, because the shared values are only the symmetric values in the matrix, the derivation still works even though it's technically incorrect (Henderson and Searle, 1979). The constrained update equation for $\mathbf{Q}$ shown in section 4.8 explicitly deals with the shared lower and upper triangles.

3.5 Update equation for a (unconstrained)

Take the derivative of Ψ with respect to $\mathbf{a}$, where $\mathbf{a}$ is a $n \times 1$ column vector. Terms not involving $\mathbf{a}$, equal 0 and drop out.

$$\begin{split}\partial\Psi/\partial\mathbf{a} = -\frac{1}{2}\sum_{t=1}^{T}\Big(-\partial(\mathrm{E}[\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\mathbf{a}])/\partial\mathbf{a} - \partial(\mathrm{E}[\mathbf{a}^\top\mathbf{R}^{-1}\boldsymbol{Y}_t])/\partial\mathbf{a} \\ + \partial(\mathrm{E}[(\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\mathbf{a}])/\partial\mathbf{a} + \partial(\mathrm{E}[\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t])/\partial\mathbf{a} + \partial(\mathrm{E}[\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{a}])/\partial\mathbf{a}\Big)\end{split} \tag{37}$$

The expectations around constants can be dropped[9]. Using relations (13) and (17) and using $\mathbf{R}^{-1} = (\mathbf{R}^{-1})^\top$, we have then

$$\begin{split}\partial\Psi/\partial\mathbf{a} = -\frac{1}{2}\sum_{t=1}^{T}\Big(-\mathrm{E}[\boldsymbol{Y}_t^\top\mathbf{R}^{-1}] - \mathrm{E}[(\mathbf{R}^{-1}\boldsymbol{Y}_t)^\top] + \mathrm{E}[(\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}] \\ + \mathrm{E}[(\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)^\top] + 2\mathbf{a}^\top\mathbf{R}^{-1}\Big)\end{split} \tag{38}$$

[9] because $\mathrm{E}_{\mathbf{XY}}(C) = C$, where C is a constant.

Pull the parameters out of the expectations, use $(\mathbf{ab})^\top = \mathbf{b}^\top\mathbf{a}^\top$ and $\mathbf{R}^{-1} = (\mathbf{R}^{-1})^\top$ where needed, and remove the $-1/2$ to get

$$\partial\Psi/\partial\mathbf{a} = \sum_{t=1}^{T}\bigg(\mathrm{E}[\boldsymbol{Y}_t]^\top\mathbf{R}^{-1} - \mathrm{E}[\boldsymbol{X}_t]^\top\mathbf{Z}^\top\mathbf{R}^{-1} - \mathbf{a}^\top\mathbf{R}^{-1}\bigg) \tag{39}$$

Set the left side to zero (a $1 \times n$ matrix of zeros), take the transpose, and cancel out $\mathbf{R}^{-1}$ by multiplying by $\mathbf{R}$, giving

$$\mathbf{0} = \sum_{t=1}^{T}\big(\mathrm{E}[\boldsymbol{Y}_t] - \mathbf{Z}\,\mathrm{E}[\boldsymbol{X}_t] - \mathbf{a}\big) = \sum_{t=1}^{T}\big(\widetilde{\mathbf{y}}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a}\big) \tag{40}$$

Solving for $\mathbf{a}$ gives us the update equation for $\mathbf{a}$:

$$\mathbf{a}_{j+1} = \frac{1}{T}\sum_{t=1}^{T}\big(\widetilde{\mathbf{y}}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t\big) \tag{41}$$

3.6 The update equation for Z (unconstrained)

Take the derivative of Ψ with respect to $\mathbf{Z}$. Terms not involving $\mathbf{Z}$, equal 0 and drop out. The expectations around terms involving only constants have been dropped.

$$\begin{aligned}
&\partial\Psi/\partial\mathbf{Z} = \text{(note } \partial\mathbf{Z} \text{ is } m \times n \text{ while } \mathbf{Z} \text{ is } n \times m\text{)}\\
&\quad -\frac{1}{2}\sum_{t=1}^{T}\bigg(- \mathrm{E}[\partial(\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{Z}]\\
&\quad - \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\boldsymbol{Y}_t)/\partial\mathbf{Z}] + \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{Z}]\\
&\quad + \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{Z}] + \mathrm{E}[\partial(\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{B}]\bigg)\\
&= -\frac{1}{2}\sum_{t=1}^{T}\bigg(- \mathrm{E}[\partial(\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{Z}]\\
&\quad - \mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Z}^\top\mathbf{R}^{-1}\boldsymbol{Y}_t)/\partial\mathbf{Z}] + \mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{Z}]\\
&\quad + \mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{Z}] + \mathrm{E}[\partial(\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{Z}]\bigg)
\end{aligned} \tag{42}$$

Using relations (14) and (16) and using $\mathbf{R}^{-1} = (\mathbf{R}^{-1})^\top$, we get

$$\begin{aligned}
\partial\Psi/\partial\mathbf{Z} = -\frac{1}{2}\sum_{t=1}^{T}\bigg(&- \mathrm{E}[\boldsymbol{X}_t\boldsymbol{Y}_t^\top\mathbf{R}^{-1}] - \mathrm{E}[\boldsymbol{X}_t\boldsymbol{Y}_t^\top\mathbf{R}^{-1}]\\
&+ 2\,\mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top\mathbf{Z}^\top\mathbf{R}^{-1}] + \mathrm{E}[\boldsymbol{X}_{t-1}\mathbf{a}^\top\mathbf{R}^{-1}] + \mathrm{E}[\boldsymbol{X}_t\mathbf{a}^\top\mathbf{R}^{-1}]\bigg)
\end{aligned} \tag{43}$$

Pulling the parameters out of the expectations and getting rid of the $-1/2$, we have

$$\partial\Psi/\partial\mathbf{Z} = \sum_{t=1}^{T}\bigg(\mathrm{E}[\boldsymbol{X}_t\boldsymbol{Y}_t^\top]\mathbf{R}^{-1} - \mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top]\mathbf{Z}^\top\mathbf{R}^{-1} - \mathrm{E}[\boldsymbol{X}_t]\mathbf{a}^\top\mathbf{R}^{-1}\bigg) \tag{44}$$

Set the left side to zero (a $m \times n$ matrix of zeros), transpose it all, and cancel out $\mathbf{R}^{-1}$ by multiplying by $\mathbf{R}$ on the left, to give

$$\begin{aligned}\mathbf{0} &= \sum_{t=1}^{T} \big(\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top] - \mathbf{Z}\,\mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top] - \mathbf{a}\,\mathrm{E}[\boldsymbol{X}_t^\top]\big)\\ &= \sum_{t=1}^{T} \big(\widetilde{\mathbf{yx}}_t - \mathbf{Z}\widetilde{\mathbf{P}}_t - \mathbf{a}\widetilde{\mathbf{x}}_t^\top\big)\end{aligned} \tag{45}$$

Solving for $\mathbf{Z}$ and noting that $\widetilde{\mathbf{P}}_t$ is invertible, gives us the new $\mathbf{Z}$:

$$\mathbf{Z}_{j+1} = \bigg(\sum_{t=1}^{T} \big(\widetilde{\mathbf{yx}}_t - \mathbf{a}\widetilde{\mathbf{x}}_t^\top\big)\bigg)\bigg(\sum_{t=1}^{T}\widetilde{\mathbf{P}}_t\bigg)^{-1} \tag{46}$$

3.7 The update equation for R (unconstrained)

Take the derivative of Ψ with respect to $\mathbf{R}$. Terms not involving $\mathbf{R}$, equal 0 and drop out. The expectations around terms involving constants have been removed.

$$\begin{aligned}\partial\Psi/\partial\mathbf{R} = -\frac{1}{2}\sum_{t=1}^{T}\bigg(&\mathrm{E}[\partial(\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\boldsymbol{Y}_t)/\partial\mathbf{R}] - \mathrm{E}[\partial(\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{R}]\\ &- \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\boldsymbol{Y}_t)/\partial\mathbf{R}] - \mathrm{E}[\partial(\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{R}]\\ &- \mathrm{E}[\partial(\mathbf{a}^\top\mathbf{R}^{-1}\boldsymbol{Y}_t)/\partial\mathbf{R}] + \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{R}]\\ &+ \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{R}] + \mathrm{E}[\partial(\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{R}]\\ &+ \partial(\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{R}\bigg) - \partial\big(\frac{T}{2}\log|\mathbf{R}|\big)/\partial\mathbf{R}\end{aligned} \tag{47}$$

We use relations (18) and (15) to do the differentiation. Notice that all the terms in the summation are of the form $\mathbf{c}^\top\mathbf{R}^{-1}\mathbf{b}$, and thus after differentiation, we group all the $\mathbf{c}^\top\mathbf{b}$ inside one set of parentheses. Also there is a minus that comes from equation (18) and cancels out the minus in front of $-1/2$.

$$\begin{aligned}\partial\Psi/\partial\mathbf{R} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{R}^{-1}\bigg(&\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top] - \mathrm{E}[\boldsymbol{Y}_t(\mathbf{Z}\boldsymbol{X}_t)^\top] - \mathrm{E}[\mathbf{Z}\boldsymbol{X}_t\boldsymbol{Y}_t^\top]\\ &- \mathrm{E}[\boldsymbol{Y}_t\mathbf{a}^\top] - \mathrm{E}[\mathbf{a}\boldsymbol{Y}_t^\top] + \mathrm{E}[\mathbf{Z}\boldsymbol{X}_t(\mathbf{Z}\boldsymbol{X}_t)^\top] + \mathrm{E}[\mathbf{Z}\boldsymbol{X}_t\mathbf{a}^\top] + \mathrm{E}[\mathbf{a}(\mathbf{Z}\boldsymbol{X}_t)^\top]\\ &+ \mathbf{a}\mathbf{a}^\top\bigg)\mathbf{R}^{-1} - \frac{T}{2}\mathbf{R}^{-1}\end{aligned} \tag{48}$$

Pulling the parameters out of the expectations and using $(\mathbf{Z}\boldsymbol{Y}_t)^\top = \boldsymbol{Y}_t^\top\mathbf{Z}^\top$, we have

$$\begin{aligned}\partial\Psi/\partial\mathbf{R} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{R}^{-1}\bigg(&\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top] - \mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top]\mathbf{Z}^\top - \mathbf{Z}\,\mathrm{E}[\boldsymbol{X}_t\boldsymbol{Y}_t^\top] - \mathrm{E}[\boldsymbol{Y}_t]\mathbf{a}^\top - \mathbf{a}\,\mathrm{E}[\boldsymbol{Y}_t^\top]\\ &+ \mathbf{Z}\,\mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top]\mathbf{Z}^\top + \mathbf{Z}\,\mathrm{E}[\boldsymbol{X}_t]\mathbf{a}^\top + \mathbf{a}\,\mathrm{E}[\boldsymbol{X}_t^\top]\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\bigg)\mathbf{R}^{-1} - \frac{T}{2}\mathbf{R}^{-1}\end{aligned} \tag{49}$$

We rewrite the partial derivative in terms of expectations:

$$\begin{aligned}\partial\Psi/\partial\mathbf{R} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{R}^{-1}\Big(&\widetilde{\mathbf{O}}_t - \widetilde{\mathbf{yx}}_t\mathbf{Z}^\top - \mathbf{Z}\widetilde{\mathbf{yx}}_t^\top - \widetilde{\mathbf{y}}_t\mathbf{a}^\top - \mathbf{a}\widetilde{\mathbf{y}}_t^\top\\ &+ \mathbf{Z}\widetilde{\mathbf{P}}_t\mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{x}}_t\mathbf{a}^\top + \mathbf{a}\widetilde{\mathbf{x}}_t^\top\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\Big)\mathbf{R}^{-1} - \frac{T}{2}\mathbf{R}^{-1}\end{aligned} \tag{50}$$

Setting this to zero (a $n \times n$ matrix of zeros), we cancel out $\mathbf{R}^{-1}$ by multiplying by $\mathbf{R}$ twice, once on the left and once on the right, and get rid of the $1/2$.

$$\begin{aligned}\mathbf{0} = \sum_{t=1}^{T}\Big(&\widetilde{\mathbf{O}}_t - \widetilde{\mathbf{yx}}_t\mathbf{Z}^\top - \mathbf{Z}\widetilde{\mathbf{yx}}_t^\top - \widetilde{\mathbf{y}}_t\mathbf{a}^\top - \mathbf{a}\widetilde{\mathbf{y}}_t^\top\\ &+ \mathbf{Z}\widetilde{\mathbf{P}}_t\mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{x}}_t\mathbf{a}^\top + \mathbf{a}\widetilde{\mathbf{x}}_t^\top\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\Big) - T\mathbf{R}\end{aligned} \tag{51}$$

We can then solve for $\mathbf{R}$, giving us the new $\mathbf{R}$ that maximizes Ψ,

$$\begin{aligned}\mathbf{R}_{j+1} = \frac{1}{T}\sum_{t=1}^{T}\Big(&\widetilde{\mathbf{O}}_t - \widetilde{\mathbf{yx}}_t\mathbf{Z}^\top - \mathbf{Z}\widetilde{\mathbf{yx}}_t^\top - \widetilde{\mathbf{y}}_t\mathbf{a}^\top - \mathbf{a}\widetilde{\mathbf{y}}_t^\top\\ &+ \mathbf{Z}\widetilde{\mathbf{P}}_t\mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{x}}_t\mathbf{a}^\top + \mathbf{a}\widetilde{\mathbf{x}}_t^\top\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\Big)\end{aligned} \tag{52}$$

As with $\mathbf{Q}$, this derivation immediately generalizes to a block diagonal matrix:

$$\mathbf{R} = \begin{bmatrix}\mathbf{R}_1 & 0 & 0\\ 0 & \mathbf{R}_2 & 0\\ 0 & 0 & \mathbf{R}_3\end{bmatrix}$$

In this case,

$$\begin{aligned}\mathbf{R}_{i,j+1} = \frac{1}{T}\sum_{t=1}^{T}\Big(&\widetilde{\mathbf{O}}_t - \widetilde{\mathbf{yx}}_t\mathbf{Z}^\top - \mathbf{Z}\widetilde{\mathbf{yx}}_t^\top - \widetilde{\mathbf{y}}_t\mathbf{a}^\top - \mathbf{a}\widetilde{\mathbf{y}}_t^\top\\ &+ \mathbf{Z}\widetilde{\mathbf{P}}_t\mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{x}}_t\mathbf{a}^\top + \mathbf{a}\widetilde{\mathbf{x}}_t^\top\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\Big)_i\end{aligned} \tag{53}$$

where the subscript i means we take the elements in the matrix in the big parentheses that are analogous to $\mathbf{R}_i$. If $\mathbf{R}_i$ is comprised of rows a to b and columns c to d of matrix $\mathbf{R}$, then we take rows a to b and columns c to d of matrix subscripted by i in equation (53).

3.8 Update equation for ξ and Λ (unconstrained), stochastic initial state

Shumway and Stoffer (2006) and Ghahramani and Hinton (1996) imply in their discussion of the EM algorithm that both ξ and Λ can be estimated (though not simultaneously). Harvey (1989), however, discusses that there are only two allowable cases:

$\boldsymbol{x}_0$ is treated as fixed ($\Lambda = 0$) and equal to the unknown parameter ξ or $\boldsymbol{x}_0$ is treated as stochastic with a known mean ξ and variance Λ. For completeness, we show here the update equation in the case of $\boldsymbol{x}_0$ stochastic with unknown mean ξ and variance Λ (a case that Harvey (1989) says is not consistent).

We proceed as before and solve for the new ξ by minimizing Ψ. Take the derivative of Ψ with respect to ξ . Terms not involving ξ, equal 0 and drop out.

$$\begin{aligned}\partial\Psi/\partial\xi = -\frac{1}{2}\big(&-\partial(\mathrm{E}[\xi^\top\Lambda^{-1}\boldsymbol{X}_0])/\partial\xi - \partial(\mathrm{E}[\boldsymbol{X}_0^\top\Lambda^{-1}\xi])/\partial\xi \\ &+\partial(\xi^\top\Lambda^{-1}\xi)/\partial\xi\big)\end{aligned} \tag{54}$$

Using relations (13) and (17) and using $\Lambda^{-1} = (\Lambda^{-1})^\top$, we have

$$\partial\Psi/\partial\xi = -\frac{1}{2}\big(-\mathrm{E}[\boldsymbol{X}_0^\top\Lambda^{-1}] - \mathrm{E}[\boldsymbol{X}_0^\top\Lambda^{-1}] + 2\xi^\top\Lambda^{-1}\big) \tag{55}$$

Pulling the parameters out of the expectations, we get

$$\partial\Psi/\partial\xi = -\frac{1}{2}\big(-2\,\mathrm{E}[\boldsymbol{X}_0^\top]\Lambda^{-1} + 2\xi^\top\Lambda^{-1}\big) \tag{56}$$

We then set the left side to zero, take the transpose, and cancel out $-1/2$ and Λ^{-1} (by noting that it is a variance-covariance matrix and is invertible).

$$\mathbf{0} = \big(\Lambda^{-1}\,\mathrm{E}[\boldsymbol{X}_0] + \Lambda^{-1}\xi\big) = (\widetilde{\mathbf{x}}_0 - \xi) \tag{57}$$

Thus,

$$\xi_{j+1} = \widetilde{\mathbf{x}}_0 \tag{58}$$

$\widetilde{\mathbf{x}}_0$ is the expected value of $\boldsymbol{X}_0$ conditioned on the data from $t = 1$ to T, which comes from the Kalman smoother recursions with initial conditions defined as $\mathrm{E}[\boldsymbol{X}_0|\boldsymbol{Y}_0 = \boldsymbol{y}_0] \equiv \xi$ and $\mathrm{var}(\boldsymbol{X}_0\boldsymbol{X}_0^\top|\boldsymbol{Y}_0 = \boldsymbol{y}_0) \equiv \Lambda$. A similar set of steps gets us to the update equation for Λ,

$$\Lambda_{j+1} = \widetilde{\mathbf{V}}_0 \tag{59}$$

$\widetilde{\mathbf{V}}_0$ is the variance of $\boldsymbol{X}_0$ conditioned on the data from $t = 1$ to T and is an output from the Kalman smoother recursions.

If the initial state is defined as at $t = 1$ instead of $t = 0$, the update equation is derived in an identical fashion and the update equation is similar:

$$\xi_{j+1} = \widetilde{\mathbf{x}}_1 \tag{60}$$

$$\Lambda_{j+1} = \widetilde{\mathbf{V}}_1 \tag{61}$$

These are output from the Kalman smoother recursions with initial conditions defined as $\mathrm{E}[\boldsymbol{X}_1|\boldsymbol{Y}_0 = \boldsymbol{y}_0] \equiv \xi$ and $\mathrm{var}(\boldsymbol{X}_1\boldsymbol{X}_1^\top|\boldsymbol{Y}_0 = \boldsymbol{y}_0) \equiv \Lambda$. Notice that the recursions are initialized slightly differently; you will see the Kalman filter and smoother equations presented with both types of initializations depending on whether the author defines the initial state at $t = 0$ or $t = 1$.

3.9 Update equation for ξ (unconstrained), fixed $\boldsymbol{x}_0$

For the case where $\boldsymbol{x}_0$ is treated as fixed, i.e. as another parameter, then there is no Λ, and we need to maximize $\partial\Psi/\partial\xi$ using the slightly different Ψ shown in equation (6). Now ξ appears in the state equation part of the likelihood.

$$\begin{aligned}\partial\Psi/\partial\xi = &-\frac{1}{2}\bigg(- \mathrm{E}[\partial(\boldsymbol{X}_1^\top \mathbf{Q}^{-1}\mathbf{B}\xi)/\partial\xi]\\ &- \mathrm{E}[\partial((\mathbf{B}\xi)^\top \mathbf{Q}^{-1}\boldsymbol{X}_1)/\partial\xi] + \mathrm{E}[\partial((\mathbf{B}\xi)^\top \mathbf{Q}^{-1}(\mathbf{B}\xi))/\partial\xi]\\ &+ \mathrm{E}[\partial((\mathbf{B}\xi)^\top \mathbf{Q}^{-1}\mathbf{u})/\partial\xi] + \mathrm{E}[\partial(\mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\xi)/\partial\xi]\bigg)\\ = &-\frac{1}{2}\bigg(- \mathrm{E}[\partial(\boldsymbol{X}_1^\top \mathbf{Q}^{-1}\mathbf{B}\xi)/\partial\xi]\\ &- \mathrm{E}[\partial(\xi^\top \mathbf{B}^\top \mathbf{Q}^{-1}\boldsymbol{X}_1)/\partial\xi] + \mathrm{E}[\partial(\xi^\top \mathbf{B}^\top \mathbf{Q}^{-1}(\mathbf{B}\xi))/\partial\xi]\\ &+ \mathrm{E}[\partial(\xi^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{u})/\partial\xi] + \mathrm{E}[\partial(\mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\xi)/\partial\xi]\bigg)\end{aligned} \tag{62}$$

After pulling the constants out of the expectations, we use relations (14) and (16) to take the derivative:

$$\begin{aligned}\partial\Psi/\partial\xi = -\frac{1}{2}\bigg(&- \mathrm{E}[\boldsymbol{X}_1]^\top \mathbf{Q}^{-1}\mathbf{B} - \mathrm{E}[\boldsymbol{X}_1]^\top \mathbf{Q}^{-1}\mathbf{B}\\ &+ 2\xi^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B} + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B} + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\bigg)\end{aligned} \tag{63}$$

This can be reduced to

$$\partial\Psi/\partial\xi = \mathrm{E}[\boldsymbol{X}_1]^\top \mathbf{Q}^{-1}\mathbf{B} - \xi^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B} - \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B} \tag{64}$$

To solve for ξ, set the left side to zero (an $m \times 1$ matrix of zeros), transpose the whole equation, and then cancel out $\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B}$ by multiplying by its inverse on the left, and solve for ξ. This step requires that this inverse exists.

$$\xi = (\mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B})^{-1}\mathbf{B}^\top \mathbf{Q}^{-1}(\mathrm{E}[\boldsymbol{X}_1] - \mathbf{u}) \tag{65}$$

Thus, in terms of the Kalman filter/smoother output the new ξ for EM iteration $j+1$ is

$$\xi_{j+1} = (\mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B})^{-1}\mathbf{B}^\top \mathbf{Q}^{-1}(\widetilde{\mathbf{x}}_1 - \mathbf{u}) \tag{66}$$

Note that using, $\widetilde{\mathbf{x}}_0$ output from the Kalman smoother would not work since $\Lambda = 0$. As a result, $\xi_{j+1} \equiv \xi_j$ in the EM algorithm, and it is impossible to move away from your starting condition for ξ.

This is conceptually similar to using a generalized least squares estimate of ξ to concentrate it out of the likelihood as discussed in Harvey (1989), section 3.4.4. However, in the context of the EM algorithm, dealing with the fixed $\boldsymbol{x}_0$ case requires nothing special; one simply takes care to use the likelihood for the case where $\boldsymbol{x}_0$ is treated as an unknown parameter (equation 6). For the other parameters, the update equations are the same whether one uses the log-likelihood equation with $\boldsymbol{x}_0$ treated as stochastic (equation 5) or fixed (equation 6).

If your MARSS model is stationary[10] and your data appear stationary, however, equation (65) probably is not what you want to use. The estimate of ξ will be the maximum-likelihood value, but it will not be drawn from the stationary distribution; instead it could be some wildly different value that happens to give the maximum-likelihood. If you are modeling the data as stationary, then you should probably assume that ξ is drawn from the stationary distribution of the $\boldsymbol{X}$'s, which is some function of your model parameters. This would mean that the model parameters would enter the part of the likelihood that involves ξ and Λ. Since you probably don't want to do that (if might start to get circular), you might try an iterative process to get decent ξ and Λ or try fixing ξ and estimating Λ (above). You can fix ξ at, say, zero, by making sure the model you fit has a stationary distribution with mean zero. You might also need to demean your data (or estimate the **a** term to account for non-zero mean data).

3.10 Update equation for ξ (unconstrained), fixed $\boldsymbol{x}_1$

In some cases, the estimate of $\boldsymbol{x}_0$ from $\boldsymbol{x}_1$ using equation 66 will be highly sensitive to small changes in the parameters. This is particularly the case for certain **B** matrices, even if they are stationary. The result is that your ξ estimate is wildly different from the data at $t = 1$. The estimates are correct given how you defined the model, just not realistic given the data. In this case, you might want to specify ξ as being the value of $\boldsymbol{x}$ at $t = 1$ instead of $t = 0$. That way, the data at $t = 1$ will constrain the estimated ξ. In this case, we treat $\boldsymbol{x}_1$ as fixed but unknown, and the variance of $\boldsymbol{X}_1$ is zero. The likelihood is then:

$$\begin{aligned}
\log \mathbf{L}(\boldsymbol{y},\boldsymbol{x};\Theta) = &-\sum_1^T \frac{1}{2}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a})^\top \mathbf{R}^{-1}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a}) - \sum_1^T \frac{1}{2}\log|\mathbf{R}| \\
&-\sum_2^T \frac{1}{2}(\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u})^\top \mathbf{Q}^{-1}(\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u}) - \sum_1^T \frac{1}{2}\log|\mathbf{Q}| \\
\boldsymbol{x}_1 \equiv \xi &
\end{aligned} \tag{67}$$

[10] meaning the $\boldsymbol{X}$'s have a stationary distribution

$$\begin{aligned}
\partial\Psi/\partial\xi = -\frac{1}{2}\bigg(&- \mathrm{E}[\partial(\boldsymbol{Y}_1^\top \mathbf{R}^{-1}\mathbf{Z}\xi)/\partial\xi] \\
&- \mathrm{E}[\partial((\mathbf{Z}\xi)^\top \mathbf{R}^{-1}\boldsymbol{Y}_1)/\partial\xi] + \mathrm{E}[\partial((\mathbf{Z}\xi)^\top \mathbf{R}^{-1}(\mathbf{Z}\xi))/\partial\xi] \\
&+ \mathrm{E}[\partial((\mathbf{Z}\xi)^\top \mathbf{R}^{-1}\mathbf{a})/\partial\xi] + \mathrm{E}[\partial(\mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z}\xi)/\partial\xi]\bigg) \\
&- \frac{1}{2}\bigg(- \mathrm{E}[\partial(\boldsymbol{X}_2^\top \mathbf{Q}^{-1}\mathbf{B}\xi)/\partial\xi] \\
&- \mathrm{E}[\partial((\mathbf{B}\xi)^\top \mathbf{Q}^{-1}\boldsymbol{X}_2)/\partial\xi] + \mathrm{E}[\partial((\mathbf{B}\xi)^\top \mathbf{Q}^{-1}(\mathbf{B}\xi))/\partial\xi] \\
&+ \mathrm{E}[\partial((\mathbf{B}\xi)^\top \mathbf{Q}^{-1}\mathbf{u})/\partial\xi] + \mathrm{E}[\partial(\mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\xi)/\partial\xi]\bigg)
\end{aligned} \tag{68}$$

Note that the second summation starts at $t = 2$ and ξ is $\boldsymbol{x}_1$ instead of $\boldsymbol{x}_0$.

After pulling the constants out of the expectations, we use relations (14) and (16) to take the derivative:

$$\begin{aligned}
\partial\Psi/\partial\xi = -\frac{1}{2}\bigg(&- \mathrm{E}[\boldsymbol{Y}_1]^\top \mathbf{R}^{-1}\mathbf{Z} - \mathrm{E}[\boldsymbol{Y}_1]^\top \mathbf{R}^{-1}\mathbf{Z} \\
&+ 2\xi^\top \mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z}\bigg) \\
&- \frac{1}{2}\bigg(- \mathrm{E}[\boldsymbol{X}_2]^\top \mathbf{Q}^{-1}\mathbf{B} - \mathrm{E}[\boldsymbol{X}_2]^\top \mathbf{Q}^{-1}\mathbf{B} \\
&+ 2\xi^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B} + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B} + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\bigg)
\end{aligned} \tag{69}$$

This can be reduced to

$$\begin{aligned}
\partial\Psi/\partial\xi &= \mathrm{E}[\boldsymbol{Y}_1]^\top \mathbf{R}^{-1}\mathbf{Z} - \xi^\top \mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} - \mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z} \\
&\quad + \mathrm{E}[\boldsymbol{X}_2]^\top \mathbf{Q}^{-1}\mathbf{B} - \xi^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B} - \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B} \\
&\quad = -\xi^\top(\mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B}) + \mathrm{E}[\boldsymbol{Y}_1]^\top \mathbf{R}^{-1}\mathbf{Z} - \mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z} \\
&\quad + \mathrm{E}[\boldsymbol{X}_2]^\top \mathbf{Q}^{-1}\mathbf{B} - \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}
\end{aligned} \tag{70}$$

To solve for ξ, set the left side to zero (an $m \times 1$ matrix of zeros), transpose the whole equation, and solve for ξ.

$$\xi = (\mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B})^{-1}(\mathbf{Z}^\top \mathbf{R}^{-1}(\mathrm{E}[\boldsymbol{Y}_1] - \mathbf{a}) + \mathbf{B}^\top \mathbf{Q}^{-1}(\mathrm{E}[\boldsymbol{X}_2] - \mathbf{u})) \tag{71}$$

Thus, when $\xi \equiv \boldsymbol{x}_1$, the new ξ for EM iteration $j+1$ is

$$\xi_{j+1} = (\mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B})^{-1}(\mathbf{Z}^\top \mathbf{R}^{-1}(\widetilde{\mathbf{y}}_1 - \mathbf{a}) + \mathbf{B}^\top \mathbf{Q}^{-1}(\widetilde{\mathbf{x}}_2 - \mathbf{u})) \tag{72}$$

4 The constrained update equations

The previous sections dealt with the case where all the elements in a parameter matrix are estimated. In this section, I deal with the case where some of the elements are constrained, for example when some matrix elements are fixed values or are linear combinations of other elements.

Let's say we have some parameter matrix $\mathbf{M}$ (here $\mathbf{M}$ could be any of the parameters in the MARSS model) where each matrix element is written as a linear model of some potentially shared values:

$$\mathbf{M} = \begin{bmatrix} a+2c+2 & 0.9 & c \\ -1.2 & a & 0 \\ 0 & 3c+1 & b \end{bmatrix}$$

Thus each i-th element in $\mathbf{M}$ can be written as $\beta_i + \beta_{a,i}a + \beta_{b,i}b + \beta_{c,i}c$, which is a linear combination of three estimated values a, b and c. The matrix $\mathbf{M}$ can be rewritten in terms of a β_i part and the part involving the $\beta_{-,j}$'s:

$$\mathbf{M} = \begin{bmatrix} 2 & 0.9 & 0 \\ -1.2 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} a+2c & 0 & c \\ 0 & a & 0 \\ 0 & 3c & b \end{bmatrix} = \mathbf{M}_{\text{fixed}} + \mathbf{M}_{\text{free}}$$

The vec function turns any matrix into a column vector by stacking the columns on top of each other. Thus,

$$\text{vec}(\mathbf{M}) = \begin{bmatrix} a+2c+2 \\ -1.2 \\ 0 \\ 0.9 \\ a \\ 3c+1 \\ c \\ 0 \\ b \end{bmatrix}$$

We can now write $\text{vec}(\mathbf{M})$ as a linear combination of $\mathbf{f} = \text{vec}(\mathbf{M}_{\text{fixed}})$ and $\mathbf{Dm} = \text{vec}(\mathbf{M}_{\text{free}})$. $\mathbf{m}$ is a $p \times 1$ column vector of the p free values, in this case $p = 3$ and the free values are a, b, c. $\mathbf{D}$ is a design matrix that translates $\mathbf{m}$ into $\text{vec}(\mathbf{M}_{\text{free}})$. For example,

$$\text{vec}(\mathbf{M}) = \begin{bmatrix} a+2c+2 \\ -1.2 \\ 0 \\ 0.9 \\ a \\ 3c+1 \\ c \\ 0 \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ -1.2 \\ 2 \\ 0.9 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \mathbf{f} + \mathbf{Dm}$$

Table 3: Kronecker and vec relations. Here $\mathbf{A}$ is $n \times m$, $\mathbf{B}$ is $m \times p$, $\mathbf{C}$ is $p \times q$. $\mathbf{a}$ is a $m \times 1$ column vector and $\mathbf{b}$ is a $p \times 1$ column vector. The symbol $\otimes$ stands for the Kronecker product: $\mathbf{A} \otimes \mathbf{C}$ is a $np \times mq$ matrix. The identity matrix, $\mathbf{I}_n$, is a $n \times n$ diagonal matrix with ones on the diagonal.

$$\mathrm{vec}(\mathbf{a}) = \mathrm{vec}(\mathbf{a}^\top) = \mathbf{a} \tag{73}$$

The vec of a column vector (or its transpose) is itself.

$$\mathrm{vec}(\mathbf{Aa}) = (\mathbf{a}^\top \otimes \mathbf{I}_n)\,\mathrm{vec}(\mathbf{A}) = \mathbf{Aa} \tag{74}$$

$\mathrm{vec}(\mathbf{Aa}) = \mathbf{Aa}$ since $\mathbf{Aa}$ is itself an $m \times 1$ column vector.

$$\mathrm{vec}(\mathbf{AB}) = (\mathbf{I}_p \otimes \mathbf{A})\,\mathrm{vec}(\mathbf{B}) = (\mathbf{B}^\top \otimes \mathbf{I}_n)\,\mathrm{vec}(\mathbf{A}) \tag{75}$$

$$\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\,\mathrm{vec}(\mathbf{B}) \tag{76}$$

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC} \otimes \mathbf{BD}) \tag{77}$$

$$\begin{gathered}(\mathbf{a} \otimes \mathbf{I}_p)\mathbf{C} = (\mathbf{a} \otimes \mathbf{C})\\ \mathbf{C}(\mathbf{a}^\top \otimes \mathbf{I}_q) = (\mathbf{a}^\top \otimes \mathbf{C})\end{gathered} \tag{78}$$

$$(\mathbf{a} \otimes \mathbf{I}_p)\mathbf{C}(\mathbf{b}^\top \otimes \mathbf{I}_q) = (\mathbf{ab}^\top \otimes \mathbf{C}) \tag{79}$$

$$\begin{gathered}(\mathbf{a} \otimes \mathbf{a}) = \mathrm{vec}(\mathbf{aa}^\top)\\ (\mathbf{a}^\top \otimes \mathbf{a}^\top) = (\mathbf{a} \otimes \mathbf{a})^\top = (\,\mathrm{vec}(\mathbf{aa}^\top))^\top\end{gathered} \tag{80}$$

There are constraints on $\mathbf{D}$. Your $\mathbf{D}$ matrix needs to describe a solvable linear set of equations. Basically it needs to be full rank (rank p where p is the number of columns in $\mathbf{D}$ or free values you are trying to estimate), so that you can estimate each of the p free values. For example, if $a+b$ always appeared together, then $a+b$ can be estimated but not a and b separately. Note, if $\mathbf{M}$ is fixed, then $\mathbf{D}$ is undefined but that is fine because in this case, there will be no update equation needed; you just use the fixed value of $\mathbf{M}$ in the algorithm.

The derivation proceeds by rewriting the likelihood as a function of $\mathrm{vec}(\mathbf{M})$, where $\mathbf{M}$ is whatever parameter matrix for which one is deriving the update equation. Then one rewrites that as a function of $\mathbf{m}$ using the relationship $\mathrm{vec}(\mathbf{M}) = \mathbf{f} + \mathbf{Dm}$. Finally, one finds the $\mathbf{m}$ that sets the derivative of Ψ with respect to $\mathbf{m}$ to zero. Conceptually, the algebraic steps in the derivation are similar to those in the unconstrained derivation. Thus, I will leave out most of the intermediate steps. The derivations require a few new matrix algebra and vec relationships; these are shown in Table 3.

4.1 The general u update equations

Since $\mathbf{u}$ is already a column vector, it can be rewritten simply as $\mathbf{u} = \mathbf{f}_u + \mathbf{D}_u\boldsymbol{\upsilon}$, where $\boldsymbol{\upsilon}$ is the column vector of estimated parameters in $\mathbf{u}$. We then solve for $\partial\Psi/\partial\boldsymbol{\upsilon}$ by replacing $\mathbf{u}$ with $\mathbf{u} = \mathbf{f}_u + \mathbf{D}_u\boldsymbol{\upsilon}$ in the expected log likelihood function. In the derivation below, the u subscripts on $\mathbf{f}$ and $\mathbf{D}$ have been left off to remove clutter.

$$\begin{gathered}
\partial\Psi/\partial\boldsymbol{\upsilon} = -\frac{1}{2}\sum_{t=1}^{T}\bigg(-\partial(\mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}(\mathbf{f}+\mathbf{D}\boldsymbol{\upsilon})])/\partial\boldsymbol{\upsilon} \\
-\partial(\mathrm{E}[(\mathbf{f}+\mathbf{D}\boldsymbol{\upsilon})^\top\mathbf{Q}^{-1}\boldsymbol{X}_t])/\partial\boldsymbol{\upsilon} + \partial(\mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}(\mathbf{f}+\mathbf{D}\boldsymbol{\upsilon})])/\partial\boldsymbol{\upsilon} \\
+\partial(\mathrm{E}[(\mathbf{f}+\mathbf{D}\boldsymbol{\upsilon})^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}])/\partial\boldsymbol{\upsilon} + \partial((\mathbf{f}+\mathbf{D}\boldsymbol{\upsilon})^\top\mathbf{Q}^{-1}(\mathbf{f}+\mathbf{D}\boldsymbol{\upsilon}))/\partial\boldsymbol{\upsilon}\bigg)
\end{gathered} \tag{81}$$

The terms involving only $\mathbf{f}$ drop out (because they don't involve $\boldsymbol{\upsilon}$). This gives

$$\begin{gathered}
\partial\Psi/\partial\boldsymbol{\upsilon} = -\frac{1}{2}\sum_{t=1}^{T}\bigg(-\partial(\mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{\upsilon}])/\partial\boldsymbol{\upsilon} - \partial(\mathrm{E}[(\mathbf{D}\boldsymbol{\upsilon})^\top\mathbf{Q}^{-1}\boldsymbol{X}_t])/\partial\boldsymbol{\upsilon} \\
+\partial(\mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{\upsilon}])/\partial\boldsymbol{\upsilon} + \partial(\mathrm{E}[(\mathbf{D}\boldsymbol{\upsilon})^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}])/\partial\boldsymbol{\upsilon} \\
+\partial(\mathbf{f}^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{\upsilon})/\partial\boldsymbol{\upsilon} + \partial((\mathbf{D}\boldsymbol{\upsilon})^\top\mathbf{Q}^{-1}\mathbf{f})/\partial\boldsymbol{\upsilon} + \partial((\mathbf{D}\boldsymbol{\upsilon})^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{\upsilon})/\partial\boldsymbol{\upsilon}\bigg)
\end{gathered} \tag{82}$$

Using the matrix differentiation relations in section 3.1, we get

$$\begin{gathered}
\partial\Psi/\partial\boldsymbol{\upsilon} = -\frac{1}{2}\sum_{t=1}^{T}\bigg(-2\,\mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{D}] + 2\,\mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{D}] \\
+2\mathbf{f}^\top\mathbf{Q}^{-1}\mathbf{D} + 2\boldsymbol{\upsilon}^\top\mathbf{D}^\top\mathbf{Q}^{-1}\mathbf{D}\bigg)
\end{gathered} \tag{83}$$

Set the left side to zero and transpose the whole equation. Then we solve for $\boldsymbol{\upsilon}$.

$$\mathbf{0} = \sum_{t=1}^{T}\Big(\mathbf{D}^\top\mathbf{Q}^{-1}(\mathrm{E}[\boldsymbol{X}_t] - \mathbf{B}\,\mathrm{E}[\boldsymbol{X}_{t-1}] - \mathbf{f}) - \mathbf{D}^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{\upsilon}\Big) \tag{84}$$

Thus,

$$T\mathbf{D}^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{\upsilon} = \mathbf{D}^\top\mathbf{Q}^{-1}\sum_{t=1}^{T}\big(\mathrm{E}[\boldsymbol{X}_t] - \mathbf{B}\,\mathrm{E}[\boldsymbol{X}_{t-1}] - \mathbf{f}\big) \tag{85}$$

Thus, the updated $\boldsymbol{\upsilon}$ is

$$\boldsymbol{\upsilon}_{j+1} = \frac{1}{T}\big(\mathbf{D}_u^\top\mathbf{Q}^{-1}\mathbf{D}_u\big)^{-1}\mathbf{D}_u^\top\mathbf{Q}^{-1}\sum_{t=1}^{T}\big(\widetilde{\mathbf{x}}_t - \mathbf{B}\widetilde{\mathbf{x}}_{t-1} - \mathbf{f}_u\big) \tag{86}$$

and

$$\mathbf{u}_{j+1} = \mathbf{f}_u + \mathbf{D}_u\boldsymbol{\upsilon}_{j+1}, \tag{87}$$

If $\mathbf{Q}$ is diagonal, this will reduce to computing the shared free elements in $\mathbf{u}$ by averaging over their values in the unconstrained $\mathbf{u}$ update matrix (equation 23.

The update equation requires that $\mathbf{D}_u^\top \mathbf{Q}^{-1} \mathbf{D}_u$ is invertible, and it will be if $\mathbf{Q}$ is a proper variance-covariance matrix (positive semi-definite) and $\mathbf{D}_u$ is full rank, as it will be if a proper variance-covariance matrix is being specified[11] and confounded elements are not being specified[12]. If $\mathbf{Q}$ has zeros on the diagonal however (a partially deterministic model), this would no longer be the case. See section 6 on the modifications to the update equation when there some of the diagonal elements of $\mathbf{Q}$ are zero.

4.2 The general a update equation

The derivation of the update equation for $\mathbf{a}$ with fixed and shared values is completely analogous to the derivation for $\mathbf{u}$. If $\mathbf{a} = \mathbf{f}_a + \mathbf{D}_a \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is a column vector of the estimated values then (with the a subscripts left of $\mathbf{D}$ and $\mathbf{f}$)

$$\boldsymbol{\alpha}_{j+1} = \frac{1}{T} \big(\mathbf{D}_a^\top \mathbf{R}^{-1} \mathbf{D}_a \big)^{-1} \mathbf{D}_a^\top \mathbf{R}^{-1} \sum_{t=1}^{T} \big(\widetilde{\mathbf{y}}_t - \mathbf{Z} \widetilde{\mathbf{x}}_t - \mathbf{f}_a \big) \tag{88}$$

The new $\mathbf{a}$ parameter is then

$$\mathbf{a}_{j+1} = \mathbf{f}_a + \mathbf{D}_a \boldsymbol{\alpha}_{j+1}, \tag{89}$$

If $\mathbf{R}$ is diagonal, this will reduce just updating the free elements in $\mathbf{a}$ using their values from the unconstrained update equation. Again $\mathbf{D}_a^\top \mathbf{R}^{-1} \mathbf{D}_a$ must be invertible; see section 6 on the modifications to the update equation when some of the diagonal elements of $\mathbf{R}$ are zero.

4.3 The general ξ update equation, stochastic initial state

When $\boldsymbol{x}_0$ is treated as stochastic with an unknown mean, the derivation of the update equation for ξ with fixed and shared values is similar to the derivation for $\mathbf{u}$ and $\mathbf{a}$. Let $\xi = \mathbf{f}_\xi + \mathbf{D}_\xi \mathbf{p}$, where $\mathbf{p}$ is a column vector of the estimated values. Take the derivative of Ψ (using equation 5) with respect to $\mathbf{p}$:

$$\partial \Psi / \partial \mathbf{p} = \big(\widetilde{\mathbf{x}}_0^\top \Lambda^{-1} - \xi^\top \Lambda^{-1} \big) \mathbf{D} \tag{90}$$

Replace ξ with $\mathbf{f} + \mathbf{D}\mathbf{p}$, set the left side to zero and transpose:

$$\mathbf{0} = \mathbf{D}^\top \big(\Lambda^{-1} \widetilde{\mathbf{x}}_0 - \Lambda^{-1} \mathbf{f} + \Lambda^{-1} \mathbf{D} \mathbf{p} \big) \tag{91}$$

Thus,

$$\mathbf{p}_{j+1} = \big(\mathbf{D}_\xi^\top \Lambda^{-1} \mathbf{D}_\xi \big)^{-1} \mathbf{D}_\xi^\top \Lambda^{-1} (\widetilde{\mathbf{x}}_0 - \mathbf{f}_\xi) \tag{92}$$

and the new ξ is then,

$$\xi_{j+1} = \mathbf{f}_\xi + \mathbf{D}_\xi \mathbf{p}_{j+1}, \tag{93}$$

When the initial state is defined as at $t = 1$, replace $\widetilde{\mathbf{x}}_0$ with $\widetilde{\mathbf{x}}_1$ in equation 92.

[11] For example, a variance-covariance matrix where all the values are equal is not valid; it's not positive semi-definite. Try taking the inverse of such a matrix; it won't work.

[12] For example, if your $\mathbf{Q}$ matrix had $a+b$ always appearing together then $a+b$ can be estimated but not a or b separately. These two parameters would be confounded.

4.4 The general ξ update equation, fixed $\boldsymbol{x}_0$

For the case where $\boldsymbol{x}_0$ is treated as fixed, i.e. as another parameter, take the derivative of Ψ using equation (6):

$$\begin{aligned}\partial\Psi/\partial\mathbf{p} = -\frac{1}{2}\Big(&- \mathrm{E}[\partial(\boldsymbol{X}_1^\top\mathbf{Q}^{-1}\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))/\partial\mathbf{p}] \\ &- \mathrm{E}[\partial((\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))^\top\mathbf{Q}^{-1}\boldsymbol{X}_1)/\partial\mathbf{p}] + \mathrm{E}[\partial((\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))^\top\mathbf{Q}^{-1}(\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p})))/\partial\mathbf{p}] \\ &+ \mathrm{E}[\partial((\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{p}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))/\partial\mathbf{p}]\Big) \\ = -\frac{1}{2}\Big(&- \mathrm{E}[\partial(\boldsymbol{X}_1^\top\mathbf{Q}^{-1}\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))/\partial\mathbf{p}] \\ &- \mathrm{E}[\partial((\mathbf{f}+\mathbf{D}\mathbf{p})^\top\mathbf{B}^\top\mathbf{Q}^{-1}\boldsymbol{X}_1)/\partial\mathbf{p}] + \mathrm{E}[\partial((\mathbf{f}+\mathbf{D}\mathbf{p})^\top\mathbf{B}^\top\mathbf{Q}^{-1}(\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p})))/\partial\mathbf{p}] \\ &+ \mathrm{E}[\partial((\mathbf{f}+\mathbf{D}\mathbf{p})^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{p}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))/\partial\mathbf{p}]\Big)\end{aligned} \tag{94}$$

After pulling the constants out of the expectations, we use relations (14) and (16) to take the derivative:

$$\begin{aligned}\partial\Psi/\partial\mathbf{p} = -\frac{1}{2}\Big(&- \mathrm{E}[\boldsymbol{X}_1]^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D} - \mathrm{E}[\boldsymbol{X}_1]^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D} \\ &+ \mathbf{f}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D} + \mathbf{f}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D} \\ &+ 2\mathbf{p}^\top\mathbf{D}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D} + \mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D} + \mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D}\Big)\end{aligned} \tag{95}$$

This can be reduced to

$$\partial\Psi/\partial\mathbf{p} = \mathrm{E}[\boldsymbol{X}_1]^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D} - \mathbf{f}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D} - \mathbf{p}^\top\mathbf{D}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D} - \mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D} \tag{96}$$

To solve for $\mathbf{p}$, set the left side to zero, transpose the whole equation, and then cancel out $\mathbf{D}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D}$ by multiplying by its inverse on the left, and solve for $\mathbf{p}$.

$$\mathbf{p} = (\mathbf{D}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{B}^\top\mathbf{Q}^{-1}(\mathrm{E}[\boldsymbol{X}_1] - \mathbf{u} - \mathbf{B}\mathbf{f}) \tag{97}$$

Thus, in terms of the Kalman filter/smoother output the new $\mathbf{p}$ for EM iteration $j+1$ is

$$\mathbf{p}_{j+1} = (\mathbf{D}_\xi^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D}_\xi)^{-1}\mathbf{D}_\xi^\top\mathbf{B}^\top\mathbf{Q}^{-1}(\widetilde{\mathbf{x}}_1 - \mathbf{u} - \mathbf{B}\mathbf{f}_\xi) \tag{98}$$

This equation requires that the inverse of $\mathbf{D}_\xi^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D}_\xi$ exists and it might not if $\mathbf{B}$ has any all zero rows/columns. In that case, defining $\xi \equiv \boldsymbol{x}_1$ might work instead (section 4.5).

4.5 The general ξ update equation, fixed $\boldsymbol{x}_1$

When the initial state is defined at $t = 1$ instead of $t = 0$, the derivation proceeds as in section 4.4 but using the likelihood in section 3.10. In terms of the Kalman smoother output the new ξ for EM iteration $j+1$ when $\xi \equiv \boldsymbol{x}_1$ is

$$\begin{aligned}\xi_{j+1} = (\mathbf{D}_\xi^\top(\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z}+\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B})\mathbf{D}_\xi)^{-1}\mathbf{D}_\xi^\top \\ (\mathbf{Z}^\top\mathbf{R}^{-1}(\widetilde{\mathbf{y}}_1-\mathbf{a}-\mathbf{f}_\xi)+\mathbf{B}^\top\mathbf{Q}^{-1}(\widetilde{\mathbf{x}}_2-\mathbf{u}-\mathbf{B}\mathbf{f}_\xi))\end{aligned} \tag{99}$$

4.6 The general B update equation

The matrix $\mathbf{B}$ is rewritten as $\mathbf{B}=\mathbf{B}_{\text{fixed}}+\mathbf{B}_{\text{free}}$, thus $\text{vec}(\mathbf{B})=\mathbf{f}_b+\mathbf{D}_b\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the $p\times 1$ column vector of the p estimated values, $\mathbf{f}_b=\text{vec}(\mathbf{B}_{\text{fixed}})$ and $\mathbf{D}_b\boldsymbol{\beta}=\text{vec}(\mathbf{B}_{\text{free}})$. Take the derivative of Ψ with respect to $\boldsymbol{\beta}$; terms in Ψ that do not involve $\mathbf{B}$ also do not involve $\boldsymbol{\beta}$ so they will equal 0 and drop out.

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\beta} = -\frac{1}{2}\sum_{t=1}^{T}\bigg(&-\text{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1})/\partial\boldsymbol{\beta}] \\ &-\text{E}[\partial((\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\boldsymbol{X}_t)/\partial\boldsymbol{\beta}]+\text{E}[\partial((\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}(\mathbf{B}\boldsymbol{X}_{t-1}))/\partial\boldsymbol{\beta}] \\ &+\text{E}[\partial((\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\boldsymbol{\beta}]+\text{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1})/\partial\boldsymbol{\beta}]\bigg)\end{aligned} \tag{100}$$

This needs to be rewritten as a function of $\boldsymbol{\beta}$ instead of $\mathbf{B}$. Note that $\mathbf{B}\boldsymbol{X}_{t-1}$ is a column vector and use relation (74) to show that:

$$\begin{aligned}\mathbf{B}\boldsymbol{X}_{t-1} = \text{vec}(\mathbf{B}\boldsymbol{X}_{t-1}) = \mathbf{K}_b\,\text{vec}(\mathbf{B}) = \mathbf{K}_b(\mathbf{f}_b+\mathbf{D}_b\boldsymbol{\beta}), \\ \text{where } \mathbf{K}_b = (\boldsymbol{X}_{t-1}^\top\otimes\mathbf{I})\end{aligned} \tag{101}$$

Thus, $\partial\Psi/\partial\boldsymbol{\beta}$ becomes (the b subscripts have been left off $\mathbf{K}$, $\mathbf{F}$ and $\mathbf{D}$):

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\beta} = -\frac{1}{2}\sum_{t=1}^{T}\bigg(&-\text{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))/\partial\boldsymbol{\beta}] \\ &-\text{E}[\partial((\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))^\top\mathbf{Q}^{-1}\boldsymbol{X}_t)/\partial\boldsymbol{\beta}] \\ &+\text{E}[\partial((\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))^\top\mathbf{Q}^{-1}\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))/\partial\boldsymbol{\beta}] \\ &+\text{E}[\partial((\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\boldsymbol{\beta}]+\text{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))/\partial\boldsymbol{\beta}]\bigg)\end{aligned} \tag{102}$$

After a bit of matrix algebra and using $\partial(\mathbf{a}^\top\mathbf{c})/\partial\mathbf{a}=\partial(\mathbf{c}^\top\mathbf{a})/\partial\mathbf{a}$, relation (13), and that partial derivatives of constants equal 0, the above can be simplified to

$$\begin{aligned}&\partial\Psi/\partial\boldsymbol{\beta} = \\ &\quad -\frac{1}{2}\sum_{t=1}^{T}\bigg(-2\,\text{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}\boldsymbol{\beta})/\partial\boldsymbol{\beta}]+2\,\text{E}[\partial((\mathbf{K}\mathbf{f})^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}\boldsymbol{\beta})/\partial\boldsymbol{\beta}] \\ &\quad +\text{E}[\partial((\mathbf{K}\mathbf{D}\boldsymbol{\beta})^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}\boldsymbol{\beta})/\partial\boldsymbol{\beta}]+2\,\text{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}\boldsymbol{\beta})/\partial\boldsymbol{\beta}]\bigg)\end{aligned} \tag{103}$$

Using relations (13) and (17), using $\mathbf{Q}^{-1} = (\mathbf{Q}^{-1})^\top$, and getting rid of the $-1/2$, we have

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\beta} = \sum_{t=1}^{T} \bigg(&\mathrm{E}[\boldsymbol{X}_t^\top \mathbf{Q}^{-1}\mathbf{K}\mathbf{D}] - \mathrm{E}[(\mathbf{K}\mathbf{f})^\top \mathbf{Q}^{-1}\mathbf{K}\mathbf{D}] \\ &- \mathrm{E}[\boldsymbol{\beta}^\top (\mathbf{K}\mathbf{D})^\top \mathbf{Q}^{-1}(\mathbf{K}\mathbf{D})] - \mathrm{E}[\mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{K}\mathbf{D}] \bigg)\end{aligned} \tag{104}$$

The left side can be set to 0 (a $1 \times p$ matrix) and the whole equation transposed, giving:

$$\begin{aligned}\mathbf{0} = \sum_{t=1}^{T} \bigg(&\mathrm{E}[(\mathbf{K}\mathbf{D})^\top \mathbf{Q}^{-1}\boldsymbol{X}_t] - \mathrm{E}[(\mathbf{K}\mathbf{D})^\top \mathbf{Q}^{-1}\mathbf{K}\mathbf{f}] \\ &- \mathrm{E}[(\mathbf{K}\mathbf{D})^\top \mathbf{Q}^{-1}(\mathbf{K}\mathbf{D})]\boldsymbol{\beta} - \mathrm{E}[(\mathbf{K}\mathbf{D})^\top \mathbf{Q}^{-1}\mathbf{u}] \bigg)\end{aligned} \tag{105}$$

Replacing $\mathbf{K}$ with $(\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I})$, we have

$$\begin{aligned}&\mathbf{0} = \\ &\sum_{t=1}^{T} \bigg(\mathrm{E}[((\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I})\mathbf{D})^\top \mathbf{Q}^{-1}\boldsymbol{X}_t] - \mathrm{E}[((\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I})\mathbf{D})^\top \mathbf{Q}^{-1}(\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I})\mathbf{f}] \\ &- \mathrm{E}[((\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I})\mathbf{D})^\top \mathbf{Q}^{-1}(\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I})\mathbf{D}]\boldsymbol{\beta} - \mathrm{E}[((\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I})\mathbf{D})^\top \mathbf{Q}^{-1}\mathbf{u}] \bigg)\end{aligned} \tag{106}$$

This looks daunting, but using relation (74) and noting that $(\mathbf{A} \otimes \mathbf{B})^\top = (\mathbf{A}^\top \otimes \mathbf{B}^\top)$, we can simplify equation (106) using the following:

$$\begin{aligned}&(\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I})\mathbf{Q}^{-1}\mathbf{u} = (\boldsymbol{X}_{t-1} \otimes \mathbf{I})\mathbf{Q}^{-1}\mathbf{u} \\ &\quad = (\boldsymbol{X}_{t-1} \otimes \mathbf{I})\,\mathrm{vec}(\mathbf{Q}^{-1}\mathbf{u}), \text{ because } \mathbf{Q}^{-1}\mathbf{u} \text{ is a column vector} \\ &\quad = \mathrm{vec}(\mathbf{Q}^{-1}\mathbf{u}(\boldsymbol{X}_{t-1})^\top), \text{ using relation (74)}\end{aligned}$$

Similarly,

$$(\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I})\mathbf{Q}^{-1}\boldsymbol{X}_t = \mathrm{vec}(\mathbf{Q}^{-1}\boldsymbol{X}_t\boldsymbol{X}_{t-1}^\top)$$

Using relation (79):

$$(\boldsymbol{X}_{t-1} \otimes \mathbf{I}_m)^\top \mathbf{Q}^{-1}(\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I}_m)\mathbf{f} = (\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top \otimes \mathbf{Q}^{-1})\mathbf{f}$$

Similarly,

$$(\boldsymbol{X}_{t-1} \otimes \mathbf{I})^\top \mathbf{Q}^{-1}(\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I})\mathbf{D}\boldsymbol{\beta} = (\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top \otimes \mathbf{Q}^{-1})\mathbf{D}\boldsymbol{\beta}$$

Using these simplifications in equation (106), we get

$$\begin{aligned}\mathbf{0} = \sum_{t=1}^{T} \bigg(&\mathrm{E}[\mathbf{D}^\top \mathrm{vec}(\mathbf{Q}^{-1}\boldsymbol{X}_t\boldsymbol{X}_{t-1}^\top)] - \mathrm{E}[\mathbf{D}^\top (\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top \otimes \mathbf{Q}^{-1})\mathbf{f}] \\ &- \mathrm{E}[\mathbf{D}^\top (\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top \otimes \mathbf{Q}^{-1})\mathbf{D}]\boldsymbol{\beta} - \mathrm{E}[\mathbf{D}^\top \mathrm{vec}(\mathbf{Q}^{-1}\mathbf{u}\boldsymbol{X}_{t-1}^\top)] \bigg)\end{aligned} \tag{107}$$

Replacing the expectations with the Kalman smoother output (section 5.1), we arrive at:

$$\begin{aligned}\mathbf{0} = \sum_{t=1}^{T}\bigg(&\mathbf{D}^\top \mathrm{vec}(\mathbf{Q}^{-1}\widetilde{\mathbf{P}}_{t,t-1}) - \mathbf{D}^\top(\widetilde{\mathbf{P}}_{t-1}\otimes\mathbf{Q}^{-1})\mathbf{f} \\ &- \mathbf{D}^\top(\widetilde{\mathbf{P}}_{t-1}\otimes\mathbf{Q}^{-1})\mathbf{D}\boldsymbol{\beta} - \mathbf{D}^\top\mathrm{vec}(\mathbf{Q}^{-1}\mathbf{u}(\widetilde{\mathbf{x}}_{t-1})^\top)\bigg)\end{aligned} \tag{108}$$

Solving for $\boldsymbol{\beta}$,

$$\begin{aligned}\boldsymbol{\beta}_{j+1} = &\bigg(\sum_{t=1}^{T}\mathbf{D}_b^\top(\widetilde{\mathbf{P}}_{t-1}\otimes\mathbf{Q}^{-1})\mathbf{D}_b\bigg)^{-1}\mathbf{D}_b^\top\bigg(\sum_{t=1}^{T}\big(\mathrm{vec}(\mathbf{Q}^{-1}\widetilde{\mathbf{P}}_{t,t-1}) \\ &- (\widetilde{\mathbf{P}}_{t-1}\otimes\mathbf{Q}^{-1})\mathbf{f}_b - \mathrm{vec}(\mathbf{Q}^{-1}\mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top)\big)\bigg)\end{aligned} \tag{109}$$

This requires that $\mathbf{D}_b^\top(\widetilde{\mathbf{P}}_{t-1}\otimes\mathbf{Q}^{-1})\mathbf{D}_b$ is invertible, and it should be since $(\widetilde{\mathbf{P}}_{t-1}\otimes\mathbf{Q}^{-1})$ is invertible[13] and $\mathbf{D}_b$ will not have any all zero columns (all zero rows are fine).

Combining $\boldsymbol{\beta}_{j+1}$ with $\mathbf{B}_{\text{fixed}}$, we arrive at the vec of the updated $\mathbf{B}$ matrix:

$$\mathrm{vec}(\mathbf{B}_{j+1}) = \mathbf{f}_b + \mathbf{D}_b\boldsymbol{\beta}_{j+1}, \tag{110}$$

When there are no fixed or shared values, $\mathbf{B}_{\text{fixed}}$ equals zero and $\mathbf{D}_b$ equals an identity matrix. Equation (109) then reduces to the unconstrained form. To see this take the vec of the unconstrained update equation for $\mathbf{B}$ and notice that $\mathbf{Q}^{-1}$ can be factored out.

4.7 The general Z update equation

The derivation of the update equation for $\mathbf{Z}$ with fixed and shared values is analogous to the derivation for $\mathbf{B}$. The matrix $\mathbf{Z}$ is rewritten as $\mathbf{Z} = \mathbf{Z}_{\text{fixed}} + \mathbf{Z}_{\text{free}}$, thus $\mathrm{vec}(\mathbf{Z}) = \mathbf{f}_z + \mathbf{D}_z\boldsymbol{\zeta}$, where $\boldsymbol{\zeta}$ is the column vector of the p estimated values, $\mathbf{f}_z = \mathrm{vec}(\mathbf{Z}_{\text{fixed}})$ and $\mathbf{D}_z\boldsymbol{\zeta} = \mathrm{vec}(\mathbf{Z}_{\text{free}})$. With the z subscript dropped off $\mathbf{D}$ and $\mathbf{f}$, the update equation for $\mathbf{Z}$ is

$$\begin{aligned}\boldsymbol{\zeta}_{j+1} = &\bigg(\sum_{t=1}^{T}(\mathbf{D}_z^\top(\widetilde{\mathbf{P}}_t\otimes\mathbf{R}^{-1})\mathbf{D}_z)\bigg)^{-1}\mathbf{D}_z^\top\bigg(\sum_{t=1}^{T}\big(\mathrm{vec}(\mathbf{R}^{-1}\widetilde{\mathbf{yx}}_t) \\ &- (\widetilde{\mathbf{P}}_t\otimes\mathbf{R}^{-1})\mathbf{f}_z - \mathrm{vec}(\mathbf{R}^{-1}\mathbf{a}\widetilde{\mathbf{x}}_t^\top)\big)\bigg)\end{aligned} \tag{111}$$

Combining $\boldsymbol{\zeta}_{j+1}$ with $\mathbf{Z}_{\text{fixed}}$, we arrive at the vec of the updated $\mathbf{Z}$ matrix:

$$\mathrm{vec}(\mathbf{Z}_{j+1}) = \mathbf{f}_z + \mathbf{D}_z\boldsymbol{\zeta}_{j+1} \tag{112}$$

This requires that $\mathbf{D}_z^\top(\widetilde{\mathbf{P}}_t\otimes\mathbf{R}^{-1})\mathbf{D}_z$ is invertible, and it should be since $(\widetilde{\mathbf{P}}_t\otimes\mathbf{R}^{-1})$ will normally be invertible[14] and $\mathbf{D}_z$ has no all zero columns.

[13] If $\mathbf{Q}$ has zeros on the diagonal, the equation needs to be altered. See section 6.

[14] If $\mathbf{R}$ has zeros on the diagonal, the equation is altered; see section 6.

4.8 The general Q update equation

A general analytical solution for fixed and shared elements in $\mathbf{Q}$ is problematic because the inverse of $\mathbf{Q}$ appears in the likelihood and because $\mathbf{Q}^{-1}$ cannot always be rewritten as a function of $\mathrm{vec}(\mathbf{Q})$. It might be an option to use numerical maximization of $\partial\Psi/\partial q_{i,j}$ where $q_{i,j}$ is a free element in $\mathbf{Q}$, but this will slow down the algorithm enormously. However, in a few important special—yet quite broad— cases, an analytical solution can be derived. The most general of these special cases is a block-symmetric matrix with optional independent fixed blocks (subsection 4.8.5). Indeed, all other cases (diagonal, block-diagonal, unconstrained, equal variance-covariance) except one (a replicated block-diagonal) are special cases of the blocked matrix with optional independent fixed blocks.

The general update equation for this case is

$$\begin{aligned}
\boldsymbol{q}_{j+1} &= \frac{1}{T}(\mathbf{D}_q^\top\mathbf{D}_q)^{-1}\mathbf{D}_q^\top\,\mathrm{vec}(\mathbf{S})\\
\mathrm{vec}(\mathbf{Q})_{j+1} &= \mathbf{f}_q + \mathbf{D}_q\boldsymbol{q}_{j+1}\\
\text{where } \mathbf{S} &= \sum_{t=1}^{T}\Big(\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}\widetilde{\mathbf{x}}_t^\top +\\
&\qquad \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\Big)
\end{aligned} \tag{113}$$

The matrices $\mathbf{f}_q$, $\mathbf{D}_q$, and $\boldsymbol{q}$ have their standard definitions. The vec of $\mathbf{Q}$ is written in the form of $\mathrm{vec}(\mathbf{Q}) = \mathbf{f}_q + \mathbf{D}_q\boldsymbol{q}$, where $\mathbf{f}_q$ is a $m^2 \times 1$ column vector of the fixed values including zero, $\mathbf{D}_q$ is the $m^2 \times p$ design matrix, and $\boldsymbol{q}$ is a column vector of the p free values. This requires that $(\mathbf{D}_q^\top\mathbf{D}_q)$, which in a valid model must be true; if is not true you have specified an invalid variance-covariance structure since the implied variance-covariance matrix will not be full-rank and thus not invertible and thus an invalid variance-covariance matrix.

Below I show how the $\mathbf{Q}$ update equation arises by working through a few of the special cases. In these derivations the q subscript is left off the $\mathbf{D}$ and $\mathbf{f}$ matrices.

4.8.1 Special case: diagonal Q matrix (with shared or unique parameters)

Let $\mathbf{Q}$ be a diagonal matrix with fixed and shared values. For example,

$$\mathbf{Q} = \begin{bmatrix} q_1 & 0 & 0 & 0 & 0\\ 0 & f_1 & 0 & 0 & 0\\ 0 & 0 & q_2 & 0 & 0\\ 0 & 0 & 0 & f_2 & 0\\ 0 & 0 & 0 & 0 & q_2 \end{bmatrix}$$

Here, f's are fixed values (constants) and q's are free parameters elements. The vec of $\mathbf{Q}^{-1}$ can be written then as $\mathrm{vec}(\mathbf{Q}^{-1}) = \mathbf{f}_q^* + \mathbf{D}_q\boldsymbol{q}^*$, where $\mathbf{f}^*$ is like $\mathbf{f}_q$ but with the corresponding i-th non-zero fixed values replaced by $1/f_i$ and $\boldsymbol{q}^*$ is a column vector of 1 over the q_i values. For the example above,

$$\boldsymbol{q}^* = \begin{bmatrix} 1/q_1\\ 1/q_2 \end{bmatrix}$$

Take the partial derivative of Ψ with respect to $\boldsymbol{q^*}$. We can do this because $\mathbf{Q}^{-1}$ is diagonal and thus each element of $\boldsymbol{q^*}$ is independent of the other elements; otherwise we would not necessarily be able to vary one element of $\boldsymbol{q^*}$ while holding the other elements constant.

$$
\begin{aligned}
\partial\Psi/\partial\boldsymbol{q^*} = -\frac{1}{2}\sum_{t=1}^{T}\partial\bigg(&\mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\boldsymbol{X}_t] - \mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}]\\
&- \mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\boldsymbol{X}_t] - \mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{u}]\\
&- \mathrm{E}[\mathbf{u}^\top\mathbf{Q}^{-1}\boldsymbol{X}_t] + \mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}]\\
&+ \mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{u}] + \mathrm{E}[\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}] + \mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{u}\bigg)/\partial\boldsymbol{q^*}\\
-\partial\big(\frac{T}{2}&\log|\mathbf{Q}|\big)/\partial\boldsymbol{q^*}
\end{aligned}
\tag{114}
$$

Using the same vec operations as in the derivations for $\mathbf{B}$ and $\mathbf{Z}$, pull $\mathbf{Q}^{-1}$ out from the middle and replace the expectations with the Kalman smoother output.[15]

$$
\begin{aligned}
\partial\Psi/\partial\boldsymbol{q^*} = -\frac{1}{2}\sum_{t=1}^{T}\partial\bigg(&\mathrm{E}[\boldsymbol{X}_t^\top\otimes\boldsymbol{X}_t^\top] - \mathrm{E}[\boldsymbol{X}_t^\top\otimes(\mathbf{B}\boldsymbol{X}_{t-1})^\top] - \mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\otimes\boldsymbol{X}_t^\top]\\
&- \mathrm{E}[\boldsymbol{X}_t^\top\otimes\mathbf{u}^\top] - \mathrm{E}[\mathbf{u}^\top\otimes\boldsymbol{X}_t^\top] + \mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\otimes(\mathbf{B}\boldsymbol{X}_{t-1})^\top]\\
&+ \mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\otimes\mathbf{u}^\top] + \mathrm{E}[\mathbf{u}^\top\otimes(\mathbf{B}\boldsymbol{X}_{t-1})^\top] + (\mathbf{u}^\top\otimes\mathbf{u}^\top)\bigg)\,\mathrm{vec}(\mathbf{Q}^{-1})/\partial\boldsymbol{q^*}\\
-\partial\bigg(\frac{T}{2}&\log|\mathbf{Q}|\bigg)/\partial\boldsymbol{q^*}\\
= -\frac{1}{2}\sum_{t=1}^{T}&\partial\big(\mathrm{vec}(\mathbf{S})^\top\big)\,\mathrm{vec}(\mathbf{Q}^{-1})/\partial\boldsymbol{q^*} + \partial\big(\frac{T}{2}\log|\mathbf{Q}^{-1}|\big)/\partial\boldsymbol{q^*}\\
\text{where } \mathbf{S} = \sum_{t=1}^{T}&\big(\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}\widetilde{\mathbf{x}}_t^\top +\\
&\mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\big)
\end{aligned}
\tag{115}
$$

Note, I have replaced $\log|\mathbf{Q}|$ with $-\log|\mathbf{Q}^{-1}|$. The determinant of a diagonal matrix is the product of its diagonal elements. Thus,

$$
\begin{aligned}
\partial\Psi/\partial\boldsymbol{q^*} = -\bigg(&\frac{1}{2}\mathrm{vec}(\mathbf{S})^\top(\mathbf{f}^* + \mathbf{D}\boldsymbol{q^*})\\
&-\frac{T}{2}(\log(f_1^*) + \log(f_2^*)...k\log(q_1^*) + l\log(q_2^*)...)\bigg)/\partial\boldsymbol{q^*}
\end{aligned}
\tag{116}
$$

where k is the number of times q_1 appears on the diagonal of $\mathbf{Q}$ and l is the number of

[15] Another, more common, way to do this is to use a "trace trick", trace($\mathbf{a}^\top\mathbf{A}\mathbf{b}$) = trace($\mathbf{A}\mathbf{b}\mathbf{a}^\top$), to pull $\mathbf{Q}^{-1}$ out.

times q_2 appears, etc. Taking the derivatives,

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{q}^* &== \frac{1}{2}\mathbf{D}^\top \operatorname{vec}(\mathbf{S}) - \frac{T}{2}(\log(f_1^*) + ...k\log(q_1^*) + l\log(q_2^*)...)/\partial\boldsymbol{q}^* \\ &= \frac{1}{2}\mathbf{D}^\top \operatorname{vec}(\mathbf{S}) - \frac{T}{2}\mathbf{D}^\top\mathbf{D}\boldsymbol{q}\end{aligned} \tag{117}$$

$\mathbf{D}^\top\mathbf{D}$ is a $p \times p$ matrix with k, l, etc. along the diagonal and thus is invertible; as usual, p is the number of free elements in $\mathbf{Q}$. Set the left side to zero (a $1 \times p$ matrix of zeros) and solve for $\boldsymbol{q}$. This gives us the update equation for $\mathbf{Q}$:

$$\begin{aligned}\boldsymbol{q}_{j+1} &= \frac{1}{T}(\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top \operatorname{vec}(\mathbf{S}) \\ \operatorname{vec}(\mathbf{Q})_{j+1} &= \mathbf{f} + \mathbf{D}\boldsymbol{q}_{j+1}\end{aligned} \tag{118}$$

where $\mathbf{S}$ is defined in equation (115) and, as usual, $\mathbf{D}$ and $\mathbf{f}$ are the parameter specific matrices. In this case, $\mathbf{D} = \mathbf{D}_q$ and $\mathbf{f} = \mathbf{f}_q$.

4.8.2 Special case: Q with one variance and one covariance

$$\mathbf{Q} = \begin{bmatrix} \alpha & \beta & \beta & \beta \\ \beta & \alpha & \beta & \beta \\ \beta & \beta & \alpha & \beta \\ \beta & \beta & \beta & \alpha \end{bmatrix} \qquad \mathbf{Q}^{-1} = \begin{bmatrix} f(\alpha,\beta) & g(\alpha,\beta) & g(\alpha,\beta) & g(\alpha,\beta) \\ g(\alpha,\beta) & f(\alpha,\beta) & g(\alpha,\beta) & g(\alpha,\beta) \\ g(\alpha,\beta) & g(\alpha,\beta) & f(\alpha,\beta) & g(\alpha,\beta) \\ g(\alpha,\beta) & g(\alpha,\beta) & g(\alpha,\beta) & f(\alpha,\beta) \end{bmatrix}$$

This is a matrix with a single shared variance parameter on the diagonal and a single shared covariance on the off-diagonals. The derivation is the same as for the diagonal case, until the step involving the differentiation of $\log|\mathbf{Q}^{-1}|$:

$$\partial\Psi/\partial\boldsymbol{q}^* = \partial\Big(-\frac{1}{2}\sum_{t=1}^{T} \big(\operatorname{vec}(\mathbf{S})^\top\big) \operatorname{vec}(\mathbf{Q}^{-1}) + \frac{T}{2}\log|\mathbf{Q}^{-1}| \Big)/\partial\boldsymbol{q}^* \tag{119}$$

It does not make sense to take the partial derivative of $\log|\mathbf{Q}^{-1}|$ with respect to $\operatorname{vec}(\mathbf{Q}^{-1})$ because many elements of $\mathbf{Q}^{-1}$ are shared so it is not possible to fix one element while varying another. Instead, we can take the partial derivative of $\log|\mathbf{Q}^{-1}|$ with respect to $g(\alpha,\beta)$ which is $\sum_{\{i,j\}\in \text{set}_g} \partial\log|\mathbf{Q}^{-1}|/\partial\boldsymbol{q}^*{}_{i,j}$. Set g is those i,j values where $\boldsymbol{q}^* = g(\alpha,\beta)$. Because $g()$ and $f()$ are different functions of both α and β, we can hold one constant while taking the partial derivative with respect to the other (well, presuming there exists some combination of α and β that would allow that). But if we have fixed values on the off-diagonal, this would not be possible. In this case (see below), we cannot hold $g()$ constant while varying $f()$ because both are only functions of α:

$$\mathbf{Q} = \begin{bmatrix} \alpha & f & f & f \\ f & \alpha & f & f \\ f & f & \alpha & f \\ f & f & f & \alpha \end{bmatrix} \qquad \mathbf{Q}^{-1} = \begin{bmatrix} f(\alpha) & g(\alpha) & g(\alpha) & g(\alpha) \\ g(\alpha) & f(\alpha) & g(\alpha) & g(\alpha) \\ g(\alpha) & g(\alpha) & f(\alpha) & g(\alpha) \\ g(\alpha) & g(\alpha) & g(\alpha) & f(\alpha) \end{bmatrix}$$

Taking the partial derivative of $\log|\mathbf{Q}^{-1}|$ with respect to $\boldsymbol{q}^* = \left[\begin{smallmatrix} f(\alpha,\beta) \\ g(\alpha,\beta) \end{smallmatrix}\right]$, we arrive at the same equation as for the diagonal matrix:

$$\partial\Psi/\partial\boldsymbol{q}^* = \frac{1}{2}\mathbf{D}^\top \operatorname{vec}(\mathbf{S}) - \frac{T}{2}\mathbf{D}^\top\mathbf{D}\boldsymbol{q} \tag{120}$$

where again $\mathbf{D}^\top\mathbf{D}$ is a $p \times p$ diagonal matrix with the number of times $f(\alpha,\beta)$ appears in element $(1,1)$ and the number of times $g(\alpha,\beta)$ appears in element $(2,2)$ of $\mathbf{D}$; $p = 2$ here since there are only 2 free parameters in $\mathbf{Q}$.

Setting to zero and solving for $\boldsymbol{q}^*$ leads to the exact same update equation as for the diagonal $\mathbf{Q}$, namely equation (118) in which $\mathbf{f}_q = 0$ since there are no fixed values.

4.8.3 Special case: a block-diagonal matrices with replicated blocks

Because these operations extend directly to block-diagonal matrices, all results for individual matrix types can be extended to a block-diagonal matrix with those types:

$$\mathbf{Q} = \begin{bmatrix} \mathbb{B}_1 & 0 & 0 \\ 0 & \mathbb{B}_2 & 0 \\ 0 & 0 & \mathbb{B}_3 \end{bmatrix}$$

where $\mathbb{B}_i$ is a matrix from any of the allowed matrix types, such as unconstrained, diagonal (with fixed or shared elements), or equal variance-covariance. Blocks can also be shared:

$$\mathbf{Q} = \begin{bmatrix} \mathbb{B}_1 & 0 & 0 \\ 0 & \mathbb{B}_2 & 0 \\ 0 & 0 & \mathbb{B}_2 \end{bmatrix}$$

but the entire block must be identical ($\mathbb{B}_2 \equiv \mathbb{B}_3$); one cannot simply share individual elements in different blocks. Either all the elements in two (or 3, or 4...) blocks are shared or none are shared.

This is ok:

$$\begin{bmatrix} c & d & d & 0 & 0 & 0 \\ d & c & d & 0 & 0 & 0 \\ d & d & c & 0 & 0 & 0 \\ 0 & 0 & 0 & c & d & d \\ 0 & 0 & 0 & d & c & d \\ 0 & 0 & 0 & d & d & c \end{bmatrix}$$

This is not ok:

$$\begin{bmatrix} c & d & d & 0 & 0 \\ d & c & d & 0 & 0 \\ d & d & c & 0 & 0 \\ 0 & 0 & 0 & c & d \\ 0 & 0 & 0 & d & c \end{bmatrix} \text{ nor } \begin{bmatrix} c & d & d & 0 & 0 & 0 \\ d & c & d & 0 & 0 & 0 \\ d & d & c & 0 & 0 & 0 \\ 0 & 0 & 0 & c & e & e \\ 0 & 0 & 0 & e & c & e \\ 0 & 0 & 0 & e & e & c \end{bmatrix}$$

The first is bad because the blocks are not identical; they need the same dimensions as well as the same values. The second is bad because again the blocks are not identical; all values must be the same.

4.8.4 Special case: a symmetric blocked matrix

The same derivation translates immediately to blocked symmetric $\mathbf{Q}$ matrices with the following form:

$$\mathbf{Q} = \begin{bmatrix} \mathbb{E}_1 & \mathbb{C}_{1,2} & \mathbb{C}_{1,3} \\ \mathbb{C}_{1,2} & \mathbb{E}_2 & \mathbb{C}_{2,3} \\ \mathbb{C}_{1,3} & \mathbb{C}_{2,3} & \mathbb{E}_3 \end{bmatrix}$$

where the $\mathbb{E}$ are as above matrices with one value on the diagonal and another on the off-diagonals (no zeros!). The $\mathbb{C}$ matrices have only one free value or are all zero. Some $\mathbb{C}$ matrices can be zero while are others are non-zero, but a individual $\mathbb{C}$ matrix cannot have a combination of free values and zero values; they have to be one or the other. Also the whole matrix must stay block symmetric. Additionally, there can be shared $\mathbb{E}$ or $\mathbb{C}$ matrices but the whole matrix needs to stay block-symmetric. Here are the forms that $\mathbb{E}$ and $\mathbb{C}$ can take:

$$\mathbb{E}_i = \begin{bmatrix} \alpha & \beta & \beta & \beta \\ \beta & \alpha & \beta & \beta \\ \beta & \beta & \alpha & \beta \\ \beta & \beta & \beta & \alpha \end{bmatrix} \qquad \mathbb{C}_i = \begin{bmatrix} \chi & \chi & \chi & \chi \\ \chi & \chi & \chi & \chi \\ \chi & \chi & \chi & \chi \\ \chi & \chi & \chi & \chi \end{bmatrix} \text{ or } \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The following are block-symmetric:

$$\begin{bmatrix} \mathbb{E}_1 & \mathbb{C}_{1,2} & \mathbb{C}_{1,3} \\ \mathbb{C}_{1,2} & \mathbb{E}_2 & \mathbb{C}_{2,3} \\ \mathbb{C}_{1,3} & \mathbb{C}_{2,3} & \mathbb{E}_3 \end{bmatrix} \text{ and } \begin{bmatrix} \mathbb{E} & \mathbb{C} & \mathbb{C} \\ \mathbb{C} & \mathbb{E} & \mathbb{C} \\ \mathbb{C} & \mathbb{C} & \mathbb{E} \end{bmatrix}$$

$$\text{and } \begin{bmatrix} \mathbb{E}_1 & \mathbb{C}_1 & \mathbb{C}_{1,2} \\ \mathbb{C}_1 & \mathbb{E}_1 & \mathbb{C}_{1,2} \\ \mathbb{C}_{1,2} & \mathbb{C}_{1,2} & \mathbb{E}_2 \end{bmatrix}$$

The following are NOT block-symmetric:

$$\begin{bmatrix} \mathbb{E}_1 & \mathbb{C}_{1,2} & 0 \\ \mathbb{C}_{1,2} & \mathbb{E}_2 & \mathbb{C}_{2,3} \\ 0 & \mathbb{C}_{2,3} & \mathbb{E}_3 \end{bmatrix} \text{ and } \begin{bmatrix} \mathbb{E}_1 & 0 & \mathbb{C}_1 \\ 0 & \mathbb{E}_1 & \mathbb{C}_2 \\ \mathbb{C}_1 & \mathbb{C}_2 & \mathbb{E}_2 \end{bmatrix} \text{ and } \begin{bmatrix} \mathbb{E}_1 & 0 & \mathbb{C}_{1,2} \\ 0 & \mathbb{E}_1 & \mathbb{C}_{1,2} \\ \mathbb{C}_{1,2} & \mathbb{C}_{1,2} & \mathbb{E}_2 \end{bmatrix}$$

$$\text{and } \begin{bmatrix} \mathbb{U}_1 & \mathbb{C}_{1,2} & \mathbb{C}_{1,3} \\ \mathbb{C}_{1,2} & \mathbb{E}_2 & \mathbb{C}_{2,3} \\ \mathbb{C}_{1,3} & \mathbb{C}_{2,3} & \mathbb{E}_3 \end{bmatrix} \text{ and } \begin{bmatrix} \mathbb{D}_1 & \mathbb{C}_{1,2} & \mathbb{C}_{1,3} \\ \mathbb{C}_{1,2} & \mathbb{E}_2 & \mathbb{C}_{2,3} \\ \mathbb{C}_{1,3} & \mathbb{C}_{2,3} & \mathbb{E}_3 \end{bmatrix}$$

In the first row, the matrices have fixed values (zeros) and free values (covariances) on the same off-diagonal row and column. That is not allowed. If there is a zero on a row or column, all other terms on the off-diagonal row and column must be also zero. In the second row, the matrix is not block-symmetric since the upper corner is an unconstrained block ($\mathbb{U}_1$) in the left matrix and diagonal block ($\mathbb{D}_1$) in the right matrix instead of a equal variance-covariance matrix ($\mathbb{E}$).

4.8.5 The general case: a block-diagonal matrix with general blocks

In it's most general form, **Q** is allowed to have a block-diagonal form where the blocks, here called $\mathbb{G}$ are any of the previous allowed cases. No shared values across $\mathbb{G}$'s; shared values are allowed within $\mathbb{G}$'s.

$$\mathbf{Q} = \begin{bmatrix} \mathbb{G}_1 & 0 & 0 \\ 0 & \mathbb{G}_2 & 0 \\ 0 & 0 & \mathbb{G}_3 \end{bmatrix}$$

The $\mathbb{G}$'s must be one of the special cases listed above: unconstrained, diagonal (with fixed or shared values), equal variance-covariance, block diagonal (with shared or unshared blocks), and block-symmetric (with shared or unshared blocks). Fixed blocks are allowed, but then the covariances with the free blocks must be zero:

$$\mathbf{Q} = \begin{bmatrix} \mathbb{F} & 0 & 0 & 0 \\ 0 & \mathbb{G}_1 & 0 & 0 \\ 0 & 0 & \mathbb{G}_2 & 0 \\ 0 & 0 & 0 & \mathbb{G}_3 \end{bmatrix}$$

Fixed blocks must have only fixed values (zero is a fixed value) but the fixed values can be different from each other. The free blocks must have only free values (zero is not a free value).

4.9 The general R update equation

The **R** update equation for blocked symmetric matrices with optional independent fixed blocks is completely analogous to the **Q** equation. Thus if **R** has the form

$$\mathbf{R} = \begin{bmatrix} \mathbb{F} & 0 & 0 & 0 \\ 0 & \mathbb{G}_1 & 0 & 0 \\ 0 & 0 & \mathbb{G}_2 & 0 \\ 0 & 0 & 0 & \mathbb{G}_3 \end{bmatrix}$$

Again the $\mathbb{G}$'s must be one of the special cases listed above: unconstrained, diagonal (with fixed or shared values), equal variance-covariance, block diagonal (with shared or unshared blocks), and block-symmetric (with shared or unshared blocks). Fixed blocks are allowed, but then the covariances with the free blocks must be zero

The update equation is

$$\begin{gathered} \boldsymbol{\rho}_{j+1} = \frac{1}{T}(\mathbf{D}_r^\top \mathbf{D}_r)^{-1}\mathbf{D}_r^\top \operatorname{vec}\left(\sum_{t=1}^{T} \mathbf{R}_{t,j+1} \right) \\ \operatorname{vec}(\mathbf{R})_{j+1} = \mathbf{f}_r + \mathbf{D}_r \boldsymbol{\rho}_{j+1} \end{gathered} \tag{121}$$

The $\mathbf{R}_{t,j+1}$ used at time step t in equation (121) is the term that appears in the summation in the unconstrained update equation with no missing values (equation 52):

$$\begin{split} \mathbf{R}_{t,j+1} = \bigg(& \widetilde{\mathbf{O}}_t - \widetilde{\mathbf{yx}}_t \mathbf{Z}^\top - \mathbf{Z}\widetilde{\mathbf{yx}}_t^\top - \widetilde{\mathbf{y}}_t \mathbf{a}^\top - \mathbf{a}\widetilde{\mathbf{y}}_t^\top \\ & + \mathbf{Z}\widetilde{\mathbf{P}}_t \mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{x}}_t \mathbf{a}^\top + \mathbf{a}\widetilde{\mathbf{x}}_t^\top \mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top \bigg) \end{split} \tag{122}$$

5 Computing the expectations in the update equations

For the update equations, we need to compute the expectations of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ and their products conditioned on 1) the observed data $\boldsymbol{Y}(1) = \boldsymbol{y}(1)$ and 2) the parameters at time t, Θ_j. This section shows how to compute these expectations. Throughout the section, I will normally leave off the conditional $\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j$ when specifying an expectation. Thus any E[] appearing without its conditional is conditioned on $\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j$. However if there are additional or different conditions those will be shown. Also all expectations are over the joint distribution of XY unless explicitly specified otherwise.

Before commencing, we need some notation for the observed and unobserved elements of the data. The $n \times 1$ vector $\boldsymbol{y}_t$ denotes the potential observations at time t. If some elements of $\boldsymbol{y}_t$ are missing, that means some elements are equal to NA (or some other missing values marker):

$$\boldsymbol{y}_t = \begin{bmatrix} y_1 \\ NA \\ y_3 \\ y_4 \\ NA \\ y_6 \end{bmatrix} \tag{123}$$

We denote the non-missing observations as $\boldsymbol{y}_t(1)$ and the missing observations as $\boldsymbol{y}_t(2)$. Similar to $\boldsymbol{y}_t$, $\boldsymbol{Y}_t$ denotes all the $\boldsymbol{Y}$ random variables at time t. The $\boldsymbol{Y}_t$'s with an observation are $\boldsymbol{Y}_t(1)$ and those without an observation are denoted $\boldsymbol{Y}_t(2)$.

Let $\Omega_t^{(1)}$ be the matrix that extracts only $\boldsymbol{Y}_t(1)$ from $\boldsymbol{Y}_t$ and $\Omega_t(2)$ be the matrix that extracts only $\boldsymbol{Y}_t(2)$. For the example above,

$$\begin{gathered} \boldsymbol{Y}_t(1) = \Omega_t^{(1)} \boldsymbol{Y}_t, \quad \Omega_t^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \\ \boldsymbol{Y}_t(2) = \Omega_t^{(2)} \boldsymbol{Y}_t, \quad \Omega_t^{(2)} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{gathered} \tag{124}$$

We will define another set of matrices that zeros out the missing or non-missing values. Let $\mathbf{I}_t^{(1)}$ denote a diagonal matrix that zeros out the $\boldsymbol{Y}_t(2)$ in $\boldsymbol{Y}_t$ and $\mathbf{I}_t^{(2)}$ denote a matrix that zeros out the $\boldsymbol{Y}_t(1)$ in $\boldsymbol{Y}_t$. For the example above,

$$\mathbf{I}_t^{(1)} = (\Omega_t^{(1)})^\top \Omega_t^{(1)} = \begin{bmatrix} 1&0&0&0&0&0\\ 0&0&0&0&0&0\\ 0&0&1&0&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&0&0\\ 0&0&0&0&0&1 \end{bmatrix} \quad \text{and}$$
$$\mathbf{I}_t^{(2)} = (\Omega_t^{(2)})^\top \Omega_t^{(2)} = \begin{bmatrix} 0&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&0&0&0&0\\ 0&0&0&0&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&0 \end{bmatrix}. \tag{125}$$

5.1 Expectations involving only $\boldsymbol{X}_t$

The Kalman smoother provides the expectations involving only $\boldsymbol{X}_t$ conditioned on all the data from time 1 to T.

$$\widetilde{\mathbf{x}}_t = \mathrm{E}[\boldsymbol{X}_t] \tag{126a}$$
$$\widetilde{\mathbf{V}}_t = \mathrm{var}[\boldsymbol{X}_t] \tag{126b}$$
$$\widetilde{\mathbf{V}}_{t,t-1} = \mathrm{cov}[\boldsymbol{X}_t,\boldsymbol{X}_{t-1}] \tag{126c}$$

From $\widetilde{\mathbf{x}}_t$, $\widetilde{\mathbf{V}}_t$, and $\widetilde{\mathbf{V}}_{t,t-1}$, we compute

$$\widetilde{\mathbf{P}}_t = \mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top] = \widetilde{\mathbf{V}}_t + \widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^\top \tag{126d}$$
$$\widetilde{\mathbf{P}}_{t,t-1} = \mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_{t-1}^\top] = \widetilde{\mathbf{V}}_{t,t-1} + \widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_{t-1}^\top \tag{126e}$$

The $\widetilde{\mathbf{P}}_t$ and $\widetilde{\mathbf{P}}_{t,t-1}$ equations arise from the computational formula for variance (equation 11). Note the smoother is different than the Kalman filter as the filter does not provide the expectations of $\boldsymbol{X}_t$ conditioned on all the data (time 1 to T) but only on the data up to time t.

The classic Kalman smoother is an algorithm to compute these expectations conditioned on no missing values in $\boldsymbol{y}$. However, the algorithm can be easily modified to give the expected values of $\boldsymbol{X}$ conditioned on the incomplete data, $\boldsymbol{Y}(1) = \boldsymbol{y}(1)$ (Shumway and Stoffer, 2006, sec. 6.4, eqn 6.78, p. 348). In this case, the usual filter and smoother equations are used with the following modifications to the parameters and data used in the equations. If the i-th element of $\boldsymbol{y}_t$ is missing, zero out the i-th rows in $\boldsymbol{y}_t$, $\mathbf{a}$ and $\mathbf{Z}$. Thus if the 2nd and 5th elements of $\boldsymbol{y}_t$ are missing,

$$\boldsymbol{y}_t = \begin{bmatrix} y_1\\0\\y_3\\y_4\\0\\y_6 \end{bmatrix}, \quad \mathbf{a}_t = \begin{bmatrix} a_1\\0\\a_3\\a_4\\0\\a_6 \end{bmatrix}, \quad \mathbf{Z}_t = \begin{bmatrix} z_{1,1} & z_{1,2} & \dots\\ 0&0&\dots\\ z_{3,1}&z_{3,2}&\dots\\ z_{4,1}&z_{4,2}&\dots\\ 0&0&\dots\\ z_{6,1}&z_{6,2}&\dots \end{bmatrix} \tag{127}$$

The $\mathbf{R}$ parameter used in the filter equations is also modified. We need to zero out the covariances between the non-missing, $\boldsymbol{y}_t(1)$, and missing, $\boldsymbol{y}_t(2)$, data. For the example above, if

$$\mathbf{R} = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} & r_{1,5} & r_{1,6} \\ r_{2,1} & r_{2,2} & r_{2,3} & r_{2,4} & r_{2,5} & r_{2,6} \\ r_{3,1} & r_{3,2} & r_{3,3} & r_{3,4} & r_{3,5} & r_{3,6} \\ r_{4,1} & r_{4,2} & r_{4,3} & r_{4,4} & r_{4,5} & r_{4,6} \\ r_{5,1} & r_{5,2} & r_{5,3} & r_{5,4} & r_{5,5} & r_{5,6} \\ r_{6,1} & r_{6,2} & r_{6,3} & r_{6,4} & r_{6,5} & r_{6,6} \end{bmatrix} \tag{128}$$

then the $\mathbf{R}$ we use at time t, will have zero covariances between the non-missing elements 1,3,4,6 and the missing elements 2,5:

$$\mathbf{R}_t = \begin{bmatrix} r_{1,1} & 0 & r_{1,3} & r_{1,4} & 0 & r_{1,6} \\ 0 & r_{2,2} & 0 & 0 & r_{2,5} & 0 \\ r_{3,1} & 0 & r_{3,3} & r_{3,4} & 0 & r_{3,6} \\ r_{4,1} & 0 & r_{4,3} & r_{4,4} & 0 & r_{4,6} \\ 0 & r_{5,2} & 0 & 0 & r_{5,5} & 0 \\ r_{6,1} & 0 & r_{6,3} & r_{6,4} & 0 & r_{6,6} \end{bmatrix} \tag{129}$$

Thus, the data and parameters used in the filter and smoother equations are

$$\begin{aligned} \boldsymbol{y}_t &= \mathbf{I}_t^{(1)} \boldsymbol{y}_t \\ \mathbf{a}_t &= \mathbf{I}_t^{(1)} \mathbf{a} \\ \mathbf{Z}_t &= \mathbf{I}_t^{(1)} \mathbf{Z} \\ \mathbf{R}_t &= \mathbf{I}_t^{(1)} \mathbf{R} \mathbf{I}_t^{(1)} + \mathbf{I}_t^{(2)} \mathbf{R} \mathbf{I}_t^{(2)} \end{aligned} \tag{130}$$

$\mathbf{a}_t$, $\mathbf{Z}_t$ and $\mathbf{R}_t$ only are used in the Kalman filter and smoother. They are not used in the EM update equations. However when coding the algorithm, it is convenient to replace the NAs (or whatever the missing values placeholder is) in $\boldsymbol{y}_t$ with zero so that there is not a problem with NAs appearing in the computations.

5.2 Expectations involving $\boldsymbol{Y}_t$

First, replace the missing values in $\boldsymbol{y}_t$ with zeros[16] and then the expectations are given by the following equations. The derivations for these equations are given in the sub-

[16] The only reason is so that in your computer code, if you use NA or NaN as the missing value marker, NA-NA=0 and 0*NA=0 rather than NA.

sections to follow.

$$\begin{aligned}
\widetilde{\mathbf{y}}_t &= \mathrm{E}[\boldsymbol{Y}_t] = \boldsymbol{y}_t - \nabla_t(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a}) && \text{(131a)}\\
\widetilde{\mathbf{O}}_t &= \mathrm{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top] = \mathbf{I}_t^{(2)}(\nabla_t\mathbf{R} + \nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_t\mathbf{Z}^\top\nabla_t^\top)\mathbf{I}_t^{(2)} + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{y}}_t^\top && \text{(131b)}\\
\widetilde{\mathbf{yx}}_t &= \mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top] = \nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_t + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_t^\top && \text{(131c)}\\
\widetilde{\mathbf{yx}}_{t,t-1} &= \mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_{t-1}^\top] = \nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_{t,t-1} + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_{t-1}^\top && \text{(131d)}\\
\text{where } \nabla_t &= \mathbf{I} - \mathbf{R}(\Omega_t^{(1)})^\top(\Omega_t^{(1)}\mathbf{R}(\Omega_t^{(1)})^\top)^{-1}\Omega_t^{(1)} && \text{(131e)}\\
\text{and } \mathbf{I}_t^{(2)} &= (\Omega_t^{(2)})^\top\Omega_t^{(2)} && \text{(131f)}
\end{aligned}$$

If $\boldsymbol{y}_t$ is all missing, $\Omega_t^{(1)}$ is a $0 \times n$ matrix, and we define $(\Omega_t^{(1)})^\top(\Omega_t^{(1)}\mathbf{R}(\Omega_t^{(1)})^\top)^{-1}\Omega_t^{(1)}$ to be a $n \times n$ matrix of zeros. If $\mathbf{R}$ is diagonal, then $\mathbf{R}(\Omega_t^{(1)})^\top(\Omega_t^{(1)}\mathbf{R}(\Omega_t^{(1)})^\top)^{-1}\Omega_t^{(1)} = \mathbf{I}_t^{(1)}$ and $\nabla_t = \mathbf{I}_t^{(2)}$. This will mean that in $\widetilde{\mathbf{y}}_t$ the $\boldsymbol{y}_t(2)$ are given by $\mathbf{Z}\widetilde{\mathbf{x}}_t + \mathbf{a}$, as expected when $\boldsymbol{y}_t(1)$ and $\boldsymbol{y}_t(2)$ are independent.

If there are zeros on the diagonal of $\mathbf{R}$ (section 6), the definition of Δ_t is changed slightly from that shown in equation 131. Let $\mho_t^{(r)}$ be the matrix that extracts the elements of $\boldsymbol{y}_t$ where $\boldsymbol{y}_t(i)$ is not missing and $\mathbf{R}(i,i)$ is not zero. Then

$$\nabla_t = \mathbf{I} - \mathbf{R}(\mho_t^{(r)})^\top(\mho_t^{(r)}\mathbf{R}(\mho_t^{(r)})^\top)^{-1}\mho_t^{(r)} \tag{132}$$

5.3 Derivation of the expected value of $\boldsymbol{Y}_t$

In the MARSS equation, the observation errors are denoted $\mathbf{v}_t$. This is a specific realization from a random variable $\mathbf{V}_t$ that is distributed multivariate normal with mean 0 and variance $\mathbf{R}$. $\mathbf{V}_t$ is not to be confused with $\widetilde{\mathbf{V}}_t$ in equation 126, which is unrelated[17] to $\mathbf{V}_t$. If there are no missing values, then we condition on $\boldsymbol{Y}_t = \boldsymbol{y}_t$ and

$$\mathrm{E}[\boldsymbol{Y}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1)] = \mathrm{E}[\boldsymbol{Y}_t|\boldsymbol{Y}_t = \boldsymbol{y}_t] = \boldsymbol{y}_t \tag{133}$$

If there are no observed values, then

$$\mathrm{E}[\boldsymbol{Y}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1)] = \mathrm{E}[\boldsymbol{Y}_t] = \mathrm{E}[\mathbf{Z}\boldsymbol{X}_t + \mathbf{a} + \mathbf{V}_t] = \mathbf{Z}\widetilde{\mathbf{x}}_t + \mathbf{a} \tag{134}$$

If only some of the $\boldsymbol{Y}_t$ are observed, then we use the conditional probability for a multivariate normal distribution (here shown for a bivariate case):

$$\text{If, } \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \tag{135}$$

Then,

$$\begin{aligned}
&(Y_1|Y_1 = y_1) = y_1, \quad \text{and}\\
&(Y_2|Y_1 = y_1) \sim \mathrm{MVN}(\bar{\mu}, \bar{\Sigma}), \quad \text{where}\\
&\qquad \bar{\mu} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_1)\\
&\qquad \bar{\Sigma} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}
\end{aligned} \tag{136}$$

[17] I apologize for the confusing notation, but $\widetilde{\mathbf{V}}_t$ and $\mathbf{v}_t$ are somewhat standard in the MARSS literature and it is standard to use a capital letter to refer to a random variable. Thus $\mathbf{V}_t$ would be the standard way to refer to the random variable associated with $\mathbf{v}_t$.

From this property, we can write down the distribution of $\boldsymbol{Y}_t$ conditioned on $\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1)$ and $\boldsymbol{X}_t = \boldsymbol{x}_t$:

$$\begin{bmatrix} \boldsymbol{Y}_t(1)|\boldsymbol{X}_t = \boldsymbol{x}_t \\ \boldsymbol{Y}_t(2)|\boldsymbol{X}_t = \boldsymbol{x}_t \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \Omega_t^{(1)}(\mathbf{Z}\boldsymbol{x}_t + \mathbf{a}) \\ \Omega_t^{(2)}(\mathbf{Z}\boldsymbol{x}_t + \mathbf{a}) \end{bmatrix}, \begin{bmatrix} \mathbf{R}_{t,11} & \mathbf{R}_{t,12} \\ \mathbf{R}_{t,21} & \mathbf{R}_{t,22} \end{bmatrix} \right) \tag{137}$$

Thus,

$$\begin{aligned} &(\boldsymbol{Y}_t(1)|\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1), \boldsymbol{X}_t = \boldsymbol{x}_t) = \Omega_t^{(1)}\boldsymbol{y}_t \quad \text{and} \\ &(\boldsymbol{Y}_t(2)|\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1), \boldsymbol{X}_t = \boldsymbol{x}_t) \sim \text{MVN}(\ddot{\mu}, \ddot{\Sigma}) \quad \text{where} \\ &\quad \ddot{\mu} = \Omega_t^{(2)}(\mathbf{Z}\boldsymbol{x}_t + \mathbf{a}) + \mathbf{R}_{t,21}(\mathbf{R}_{t,11})^{-1}\Omega_t^{(1)}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a}) \\ &\quad \ddot{\Sigma} = \mathbf{R}_{t,22} - \mathbf{R}_{t,21}(\mathbf{R}_{t,11})^{-1}\mathbf{R}_{t,12} \end{aligned} \tag{138}$$

Note that since we are conditioning on $\boldsymbol{X}_t = \boldsymbol{x}_t$, we can replace $\boldsymbol{Y}$ by $\boldsymbol{Y}_t$ in the conditional:

$$\text{E}[\boldsymbol{Y}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \boldsymbol{X}_t = \boldsymbol{x}_t] = \text{E}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1), \boldsymbol{X}_t = \boldsymbol{x}_t].$$

From this and the distributions in equation (138), we can write down $\widetilde{\mathbf{y}}_t = \text{E}[\boldsymbol{Y}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j]$:

$$\begin{aligned} \widetilde{\mathbf{y}}_t &= \text{E}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1)] \\ &= \int_{\boldsymbol{x}_t}\int_{\boldsymbol{y}_t} \boldsymbol{y}_t f(\boldsymbol{y}_t|\boldsymbol{y}_t(1), \boldsymbol{x}_t) d\boldsymbol{y}_t f(\boldsymbol{x}_t) d\boldsymbol{x}_t \\ &= \text{E}_X[\text{E}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1), \boldsymbol{X}_t = \boldsymbol{x}_t]] \\ &= \text{E}_X[\boldsymbol{y}_t - \nabla_t(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{X}_t - \mathbf{a})] \\ &= \boldsymbol{y}_t - \nabla_t(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a}) \\ \text{where } \nabla_t &= \mathbf{I} - \mathbf{R}(\Omega_t^{(1)})^\top(\mathbf{R}_{t,11})^{-1}\Omega_t^{(1)} \end{aligned} \tag{139}$$

$(\Omega_t^{(1)})^\top(\mathbf{R}_{t,11})^{-1}\Omega_t^{(1)}$ is a $n \times n$ matrix with 0s in the non-(11) positions. If the k-th element of $\boldsymbol{y}_t$ is observed, then k-th row and column of ∇_t will be zero. Thus if there are no missing values at time t, $\nabla_t = \mathbf{I} - \mathbf{I} = 0$. If there are no observed values at time t, ∇_t will reduce to $\mathbf{I}$.

5.4 Derivation of the expected value of $\boldsymbol{Y}_t\boldsymbol{Y}_t^\top$

The following outlines a[18] derivation. If there are no missing values, then we condition on $\boldsymbol{Y}_t = \boldsymbol{y}_t$ and

$$\begin{aligned} \text{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}(1) = \boldsymbol{y}(1)] &= \text{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}_t = \boldsymbol{y}_t] \\ &= \boldsymbol{y}_t\boldsymbol{y}_t^\top. \end{aligned} \tag{140}$$

[18] The following derivations are painfully ugly, but appear to work. There are surely more elegant ways to do this; at least, there must be more elegant notations.

If there are no observed values at time t, then

$$\begin{aligned}
&\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top]\\
&\quad= \mathrm{var}[\mathbf{Z}\boldsymbol{X}_t+\mathbf{a}+\mathbf{V}_t]+\mathrm{E}[\mathbf{Z}\boldsymbol{X}_t+\mathbf{a}+\mathbf{V}_t]\,\mathrm{E}[\mathbf{Z}\boldsymbol{X}_t+\mathbf{a}+\mathbf{V}_t]^\top\\
&\quad= \mathrm{var}[\mathbf{V}_t]+\mathrm{var}[\mathbf{Z}\boldsymbol{X}_t]+(\mathrm{E}[\mathbf{Z}\boldsymbol{X}_t+\mathbf{a}]+\mathrm{E}[\mathbf{V}_t])(\mathrm{E}[\mathbf{Z}\boldsymbol{X}_t+\mathbf{a}]+\mathrm{E}[\mathbf{V}_t])^\top\\
&\quad= \mathbf{R}+\mathbf{Z}\widetilde{\mathbf{V}}_t\mathbf{Z}^\top+(\mathbf{Z}\widetilde{\mathbf{x}}_t+\mathbf{a})(\mathbf{Z}\widetilde{\mathbf{x}}_t+\mathbf{a})^\top
\end{aligned}\tag{141}$$

When only some of the $\boldsymbol{Y}_t$ are observed, we use again the conditional probability of a multivariate normal (equation 135). From this property, we know that

$$\begin{aligned}
&\mathrm{var}_{Y|x}[\boldsymbol{Y}_t(2)\boldsymbol{Y}_t(2)^\top|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]=\mathbf{R}_{t,22}-\mathbf{R}_{t,21}(\mathbf{R}_{t,11})^{-1}\mathbf{R}_{t,12},\\
&\mathrm{var}_{Y|x}[\boldsymbol{Y}_t(1)|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]=0\\
\text{and }&\mathrm{cov}_{Y|x}[\boldsymbol{Y}_t(1),\boldsymbol{Y}_t(2)|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]=0
\end{aligned}$$

$$\begin{aligned}
\text{Thus }&\mathrm{var}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]\\
&\quad=(\Omega_t^{(2)})^\top(\mathbf{R}_{t,22}-\mathbf{R}_{t,21}(\mathbf{R}_{t,11})^{-1}\mathbf{R}_{t,12})\Omega_t^{(2)}\\
&\quad=(\Omega_t^{(2)})^\top(\Omega_t^{(2)}\mathbf{R}(\Omega_t^{(2)})^\top-\Omega_t^{(2)}\mathbf{R}(\Omega_t^{(1)})^\top(\mathbf{R}_{t,11})^{-1}\Omega_t^{(1)}\mathbf{R}(\Omega_t^{(2)})^\top)\Omega_t^{(2)}\\
&\quad=\mathbf{I}_t^{(2)}(\mathbf{R}-\mathbf{R}(\Omega_t^{(1)})^\top(\mathbf{R}_{t,11})^{-1}\Omega_t^{(1)}\mathbf{R})\mathbf{I}_t^{(2)}\\
&\quad=\mathbf{I}_t^{(2)}\nabla_t\mathbf{R}\mathbf{I}_t^{(2)}
\end{aligned}\tag{142}$$

The $\mathbf{I}_t^{(2)}$ bracketing both sides is zero-ing out the rows and columns corresponding to the $\boldsymbol{y}_t(1)$ values.

Now we can compute the $\mathrm{E}_{XY}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]$. The subscripts are added to the E to emphasize that we are breaking the multivariate expectation into an inner and outer expectation.

$$\begin{aligned}
\widetilde{\mathbf{O}}_t&=\mathrm{E}_{XY}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]=\mathrm{E}_X[\mathrm{E}_{Y|x}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]]\\
&=\mathrm{E}_X\big[\mathrm{var}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]\\
&\quad+\mathrm{E}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]\,\mathrm{E}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]^\top\big]\\
&=\mathrm{E}_X[\mathbf{I}_t^{(2)}\nabla_t\mathbf{R}\mathbf{I}_t^{(2)}]+\mathrm{E}_X[(\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a}))(\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a}))^\top]\\
&=\mathbf{I}_t^{(2)}\nabla_t\mathbf{R}\mathbf{I}_t^{(2)}+\mathrm{var}_X\big[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a})\big]\\
&\quad+\mathrm{E}_X[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a})]\,\mathrm{E}_X[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a})]^\top\\
&=\mathbf{I}_t^{(2)}\nabla_t\mathbf{R}\mathbf{I}_t^{(2)}+\mathbf{I}_t^{(2)}\nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_t\mathbf{Z}^\top\nabla_t^\top\mathbf{I}_t^{(2)}+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{y}}_t^\top
\end{aligned}\tag{143}$$

Thus,

$$\widetilde{\mathbf{O}}_t=\mathbf{I}_t^{(2)}(\nabla_t\mathbf{R}+\nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_t\mathbf{Z}^\top\nabla_t^\top)\mathbf{I}_t^{(2)}+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{y}}_t^\top\tag{144}$$

5.5 Derivation of the expected value of $\boldsymbol{Y}_t\boldsymbol{X}_t^\top$

If there are no missing values, then we condition on $\boldsymbol{Y}_t = \boldsymbol{y}_t$ and

$$\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)] = \boldsymbol{y}_t\,\mathrm{E}[\boldsymbol{X}_t^\top] = \boldsymbol{y}_t\widetilde{\mathbf{x}}_t^\top \tag{145}$$

If there are no observed values at time t, then

$$\begin{aligned}
&\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]\\
&\quad= \mathrm{E}[(\mathbf{Z}\boldsymbol{X}_t+\mathbf{a}+\mathbf{V}_t)\boldsymbol{X}_t^\top]\\
&\quad= \mathrm{E}[\mathbf{Z}\boldsymbol{X}_t\boldsymbol{X}_t^\top+\mathbf{a}\boldsymbol{X}_t^\top+\mathbf{V}_t\boldsymbol{X}_t^\top]\\
&\quad= \mathbf{Z}\widetilde{\mathbf{P}}_t+\mathbf{a}\widetilde{\mathbf{x}}_t^\top+\mathrm{cov}[\mathbf{V}_t,\boldsymbol{X}_t]+\mathrm{E}[\mathbf{V}_t]\,\mathrm{E}[\boldsymbol{X}_t]^\top\\
&\quad= \mathbf{Z}\widetilde{\mathbf{P}}_t+\mathbf{a}\widetilde{\mathbf{x}}_t^\top
\end{aligned} \tag{146}$$

Note that $\mathbf{V}_t$ and $\boldsymbol{X}_t$ are independent (equation 1). $\mathrm{E}[\mathbf{V}_t]=0$ and $\mathrm{cov}[\mathbf{V}_t,\boldsymbol{X}_t]=0$.

Now we can compute the $\mathrm{E}_{XY}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}_(1)=\boldsymbol{y}(1)]$.

$$\begin{aligned}
\widetilde{\mathbf{yx}}_t &= \mathrm{E}_{XY}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]\\
&= \mathrm{cov}[\boldsymbol{Y}_t,\boldsymbol{X}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1)]+\mathrm{E}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]\,\mathrm{E}_{XY}[\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]^\top\\
&= \mathrm{cov}[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a})+\mathbf{V}_t^*,\boldsymbol{X}_t]+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_t^\top\\
&= \mathrm{cov}[\boldsymbol{y}_t,\boldsymbol{X}_t]-\mathrm{cov}[\nabla_t\boldsymbol{y}_t,\boldsymbol{X}_t]+\mathrm{cov}[\nabla_t\mathbf{Z}\boldsymbol{X}_t,\boldsymbol{X}_t]+\mathrm{cov}[\nabla_t\mathbf{a},\boldsymbol{X}_t]\\
&\quad+\mathrm{cov}[\mathbf{V}_t^*,\boldsymbol{X}_t]+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_t^\top\\
&= 0-0+\nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_t+0+0+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_t^\top\\
&= \nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_t+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_t^\top
\end{aligned} \tag{147}$$

This uses the computational formula for covariance: $\mathrm{E}[\boldsymbol{Y}\boldsymbol{X}^\top]=\mathrm{cov}[\boldsymbol{Y},\boldsymbol{X}]+\mathrm{E}[\boldsymbol{Y}]\,\mathrm{E}[\boldsymbol{X}]^\top$. $\mathbf{V}_t^*$ is a random variable with mean 0 and variance $\mathbf{R}_{t,22}-\mathbf{R}_{t,21}(\mathbf{R}_{t,11})^{-1}\mathbf{R}_{t,12}$ from equation (138). $\mathbf{V}_t^*$ and $\boldsymbol{X}_t$ are independent of each other, thus $\mathrm{cov}[\mathbf{V}_t^*,\boldsymbol{X}_t^\top]=0$.

5.6 Derivation of the expected value of $\boldsymbol{Y}_t\boldsymbol{X}_{t-1}^\top$

The derivation of $\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_{t-1}^\top]$ is similar to the derivation of $\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_{t-1}^\top]$:

$$\begin{aligned}
\widetilde{\mathbf{yx}}_t &= \mathrm{E}_{XY}[\boldsymbol{Y}_t\boldsymbol{X}_{t-1}^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]\\
&= \mathrm{cov}[\boldsymbol{Y}_t,\boldsymbol{X}_{t-1}|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1)]+\mathrm{E}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]\,\mathrm{E}_{XY}[\boldsymbol{X}_{t-1}^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]^\top\\
&= \mathrm{cov}[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a})+\mathbf{V}_t^*,\boldsymbol{X}_{t-1}]+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_{t-1}^\top\\
&= \mathrm{cov}[\boldsymbol{y}_t,\boldsymbol{X}_{t-1}]-\mathrm{cov}[\nabla_t\boldsymbol{y}_t,\boldsymbol{X}_{t-1}]+\mathrm{cov}[\nabla_t\mathbf{Z}\boldsymbol{X}_t,\boldsymbol{X}_{t-1}]\\
&\quad+\mathrm{cov}[\nabla_t\mathbf{a},\boldsymbol{X}_{t-1}]+\mathrm{cov}[\mathbf{V}_t^*,\boldsymbol{X}_{t-1}]+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_{t-1}^\top\\
&= 0-0+\nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_{t,t-1}+0+0+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_{t-1}^\top\\
&= \nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_{t,t-1}+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_{t-1}^\top
\end{aligned} \tag{148}$$

6 Degenerate variance modifications

It is possible that the model has deterministic and probabilistic elements; mathematically this means that one or the other of **R** or **Q** have zeros on the diagonal in which case some of the observation or state processes are deterministic. Assuming the model is solvable (one solution and not over-determined), we can modify the Kalman smoother and EM algorithm to handle models with deterministic elements. The motivation behind the degenerate variance modification is that we want to use one set of EM update equations for all models in the MARSS class—regardless of whether they are partially or fully degenerate. The notation here is painful, but the actual math is not difficult to implement.

As an example of a solvable versus unsolvable model, consider the following. If

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & a \neq 0 & 0 & 0 \\ 0 & 0 & b \neq 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{149}$$

then following are bad versus ok **Z** matrices.

$$\mathbf{Z}_{\text{bad}} = \begin{bmatrix} c & d & 0 \\ z(2,1) & z(2,2) & z(2,3) \\ z(3,1) & z(3,1) & z(3,1) \\ c & d & 0 \end{bmatrix}, \quad \mathbf{Z}_{\text{ok}} = \begin{bmatrix} c & 0 & 0 \\ z(2,1) & z(2,2) & z(2,3) \\ z(3,1) & z(3,1) & z(3,1) \\ c & d \neq 0 & 0 \end{bmatrix} \tag{150}$$

Because $y_t(1)$ and $y_t(4)$ have zero observation variance, the first **Z** reduces to this for $x_t(1)$ and $x_t(2)$:

$$\begin{bmatrix} y_t(1) \\ y_t(4) \end{bmatrix} = \begin{bmatrix} cx_t(1) + dx_t(2) \\ cx_t(1) + dx_t(2) \end{bmatrix} \tag{151}$$

and since $y_t(1) \neq y_t(4)$, potentially, that is not solvable. The second **Z** reduces to

$$\begin{bmatrix} y_t(1) \\ y_t(4) \end{bmatrix} = \begin{bmatrix} cx_t(1) \\ cx_t(1) + dx_t(4) \end{bmatrix} \tag{152}$$

and that is solvable for any $y_t(1)$ and $y_t(4)$ combination. Notice that in the latter case, $x_t(1)$ and $x_t(2)$ are fully specified by $y_t(1)$ and $y_t(4)$. This property will be used below to deal with numerical errors that crop up when diagonal elements of **R** are equal to zero.

6.1 Kalman filter and smoother modifications

In principle, when one of the **Q** or **R** variances is zero[19], the standard Kalman filter/smoother equations would still work and provide the correct state outputs and likelihood. In practice however errors will be generated if one passes a variance matrix with zeros on the diagonal because under certain situations, one of the matrix inverses

[19] The corresponding covariances will also be zero.

will involve a matrix with a zero on the diagonal in the Kalman filter/smoother equations and this will lead to an the computer code throwing an error.

When $\mathbf{R}$ has zeros on the diagonal, problems arise in the Kalman update part of the Kalman filter. The Kalman gain[20] is

$$\mathbf{K}_t = \mathbf{V}_t^{t-1}\mathbf{Z}_t^\top(\mathbf{Z}_t\mathbf{V}_t^{t-1}\mathbf{Z}_t^\top + \mathbf{R}_t)^{-1} \tag{153}$$

Here, $\mathbf{Z}_t$ is the missing values modified $\mathbf{Z}$ matrix with the i-th rows zero-ed out if the i-th element of $\boldsymbol{y}_t$ is missing (section 5.1, equation 127). Thus if the i-th element of $\boldsymbol{y}_t$ is missing and the (i,i) element of $\mathbf{R}$ is zero, the (i,i) element of $(\mathbf{Z}_t\mathbf{V}_t^{t-1}\mathbf{Z}_t^\top + \mathbf{R}_t)$ will be zero also and one cannot take its inverse. In addition, if the initial value $\boldsymbol{x}_1$ is treated as fixed but unknown then $\mathbf{V}_1^0$ will be a $m \times m$ matrix of zeros. Again in this situation $(\mathbf{Z}_t\mathbf{V}_t^{t-1}\mathbf{Z}_t^\top + \mathbf{R}_t)$ will have zeros at any (i,i) elements where $\mathbf{R}$ is also zero.

The first case, where zeros on the diagonal arise due to missing values in the data, can be solved using the matrix which pulls out the rows and columns corresponding to the non-missing values ($\Omega_t^{(1)}$). Replace $\big(\mathbf{Z}_t\mathbf{V}_t^{t-1}\mathbf{Z}_t^\top + \mathbf{R}_t\big)^{-1}$ in equation (153) with

$$(\Omega_t^{(1)})^\top\big(\Omega_t^{(1)}(\mathbf{Z}_t\mathbf{V}_t^{t-1}\mathbf{Z}_t^\top + \mathbf{R}_t)(\Omega_t^{(1)})^\top\big)^{-1}\Omega_t^{(1)} \tag{154}$$

Wrapping in $\Omega_t^{(1)}(\Omega_t^{(1)})^\top$ gets rid of all the zero rows/columns in $\mathbf{Z}_t\mathbf{V}_t^{t-1}\mathbf{Z}_t^\top + \mathbf{R}_t$, and the matrix is reassembled with the zero rows/columns reinserted by wrapping in $(\Omega_t^{(1)})^\top\Omega_t^{(1)}$. This works because $\mathbf{R}_t$ is the missing values modified $\mathbf{R}$ (section 1.3) and is block diagonal across the i and non-i rows/columns, and $\mathbf{Z}_t$ has the i-columns zero-ed out. Thus removing the i columns and rows before taking the inverse has no effect on the product $\mathbf{Z}_t(...)^{-1}$. When $\mathbf{V}_1^0 = \mathbf{0}$, set $\mathbf{K}_1 = \mathbf{0}$ without computing the inverse (see equation 153 where $\mathbf{V}_1^0$ appears on the left).

There is also a numerical issue to deal with. When the (i,i) elements of $\mathbf{R}$ are zero, some of the elements of $\boldsymbol{x}_t$ may be completely specified (fully known) given $\boldsymbol{y}_t$. Let's call these fully known elements of $\boldsymbol{x}_t$, the k-th elements. In this case, the k-th row and column of $\mathbf{V}_t^t$ must be zero because given $y_t(i)$, $x_t(k)$ is known (is fixed) and its variance, $\mathbf{V}_t^t(k,k)$, is zero. Because $\mathbf{K}_t$ is computed using a numerical estimate of the inverse, the standard $\mathbf{V}_t^t$ update equation (which uses $\mathbf{K}_t$) will cause these elements to be close to zero but not precisely zero, and they may even be slightly negative on the diagonal. This will cause serious problems when the Kalman filter output is passed on to the EM algorithm. Thus after $\mathbf{V}_t^t$ is computed using the normal Kalman update equation, we will want to explicitly zero out the k rows and columns in the filter.

When $\mathbf{Q}$ has zeros on the diagonal, then we might also have similar numerical errors in $\mathbf{J}$ in the Kalman smoother. The $\mathbf{J}$ equation[21] is

$$\begin{aligned}\mathbf{J}_t &= \mathbf{V}_{t-1}^{t-1}\mathbf{B}^\top(\mathbf{V}_t^{t-1})^{-1}\\ &\text{where } \mathbf{V}_t^{t-1} = \mathbf{B}\mathbf{V}_{t-1}^{t-1}\mathbf{B}^\top + \mathbf{Q}\end{aligned} \tag{155}$$

[20] Refer to Shumway and Stoffer, e.g., for the Kalman filter equations. I am skipping over that to just show the changes to the recursion equations.

[21] Again, refer to Shumway and Stoffer for the Kalman filter recursions.

If there are zeros on the diagonals of (Λ and/or $\mathbf{B}$) and $\mathbf{Q}$ and these zeros line up, then if the $\mathbf{B}^{(0)}$ and $\mathbf{B}^{(1)}$ elements in $\mathbf{B}$ are blocks[22], there will be zeros on the diagonal of $\mathbf{V}_t^t$. Thus there will be zeros on the diagonal of $\mathbf{V}_t^{t-1}$ and it cannot be inverted. In this case, the corresponding elements of $\mathbf{V}_t^T$ need to be zero since what's happening is that those elements are deterministic and thus have 0 variance.

We want to catch these zero variances in $\mathbf{V}_t^{t-1}$ so that we can take the inverse. Note that this can only happen when there are zeros on the diagonal of $\mathbf{Q}$ since $\mathbf{B}\mathbf{V}_{t-1}^{t-1}\mathbf{B}^\top$ can never be negative on the diagonal since $\mathbf{B}\mathbf{B}^\top$ must be positive-definite and so is $\mathbf{V}_{t-1}^{t-1}$. The basic idea is the same as above. We replace $(\mathbf{V}_t^{t-1})^{-1}$ with:

$$(\Omega_{Vt}^+)^\top \big(\Omega_{Vt}^+ (\mathbf{V}_t^{t-1})(\Omega_{Vt}^+)^\top\big)^{-1}\Omega_{Vt}^+ \tag{156}$$

where Ω_{Vt}^+ is a matrix that removes all the positive $\mathbf{V}_t^{t-1}$ rows analogous to $\Omega_t^{(1)}$.

6.2 EM algorithm modifications

[1/25/2012 note. This whole section is under construction and very rough, but I'm posting as is to get version 2.9 up.]

The constrained update equations for $\mathbf{Q}$ and $\mathbf{R}$ (either diagonal w/o missing values or non-diagonal with no missing values) work fine because they deal with fixed values (in this case, zeros) and the derivation does not involve any inverses of non-invertible matrices. However if $\mathbf{R}$ is non-diagonal and there are missing values, then the $\mathbf{R}$ update equation involves $\widetilde{\mathbf{y}}_t$, and that will involve the inverse of $\mathbf{R}_{11}$ (section 5.2), which might have zeros on the diagonal. In that case, use the ∇_t modification that deals with zeros on the diagonal of $\mathbf{R}$ (equation 132).

6.2.1 Modified likelihood for partially deterministic models

Let $\mathbf{R}^+$ be the sub-setted positive $\mathbf{R}$ matrix. For example, if

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & .2 \\ 0 & 0 & 0 \\ .2 & 0 & 1 \end{bmatrix}, \quad \text{then} \quad \mathbf{R}^+ = \begin{bmatrix} 1 & .2 \\ .2 & 1 \end{bmatrix}. \tag{157}$$

Let Ω_r^+ be a $p \times n$ matrix that extracts the p non-zero rows from $\mathbf{R}$, and can extract $\mathbf{R}^+$ from $\mathbf{R}$. The diagonal matrix $(\Omega_r^+)^\top\Omega_r^+ \equiv \mathbf{I}_r^+$ zero's out the zero row in $\mathbf{R}$ (and any $n \times 1$ row vector. For the example above,

$$\begin{aligned} &\mathbf{R}^+ = \Omega_r^+\mathbf{R}(\Omega_r^+)^\top \\ &\boldsymbol{y}_t^+ = \Omega_r^+\boldsymbol{y}_t \\ &\Omega_r^+ = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{I}_r^+ = (\Omega_r^+)^\top\Omega_r^+ = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{158}$$

[22] This means the following. Let the rows where the diagonal elements in $\mathbf{Q}$ equal zero be denoted i and the the rows where there are non-zero diagonals be denoted j. The $\mathbf{B}^{(0)}$ elements are the $\mathbf{B}$ elements where both row and column are in i. The $\mathbf{B}^{(1)}$ elements are the $\mathbf{B}$ elements where both row and column are in j. If the $\mathbf{B}^{(0)}$ and $\mathbf{B}^{(1)}$ elements in $\mathbf{B}$ are blocks, this means all the $\mathbf{B}_{i,j}$ are 0; no deterministic components interact with the stochastic components.

Let $\Omega_r^{(0)}$ be a $(n-p)\times n$ matrix that extracts the $n-p$ zero rows from $\mathbf{R}$. For the example above,

$$\begin{gathered}\mathbf{R}^{(0)} = \Omega_r^{(0)}\mathbf{R}(\Omega_r^{(0)})^\top \\ \boldsymbol{y}_t^{(0)} = \Omega_r^{(0)}\boldsymbol{y}_t \\ \Omega_r^{(0)} = \begin{bmatrix}0 & 1 & 0\end{bmatrix} \qquad \mathbf{I}_r^{(0)} = (\Omega_r^{(0)})^\top\Omega_r^{(0)} = \begin{bmatrix}0&0&0\\0&1&0\\0&0&0\end{bmatrix}\end{gathered} \tag{159}$$

Similarly, Ω_q^+ extracts the non-zero rows from $\mathbf{Q}$ and $\Omega_q^{(0)}$ extracts the zero rows.

Using these definitions, we can rewrite the state process part of the MARSS model by separating out the deterministic parts ($\mathbf{Q}=0$):

$$\begin{aligned}\boldsymbol{x}_t^{(0)} &= \Omega_q^{(0)}\boldsymbol{x}_t = \Omega_q^{(0)}\mathbf{B}\boldsymbol{x}_{t-1} + \Omega_q^{(0)}\mathbf{u} \\ \boldsymbol{x}_t^{+} &= \Omega_q^{+}\boldsymbol{x}_t = \Omega_q^{+}\mathbf{B}\boldsymbol{x}_{t-1} + \Omega_q^{+}\mathbf{u} + \mathbf{w}_t^+ \\ \mathbf{w}_t^+ &\sim \text{MVN}(0,\mathbf{Q}^+) \\ \boldsymbol{x}_0 &\sim \text{MVN}(\xi,\Lambda)\end{aligned} \tag{160}$$

Similarly, we can rewrite the observation process part of the MARSS model by separating out the parts with $\mathbf{R}=0$:

$$\begin{aligned}\boldsymbol{y}_t^{(0)} &= \Omega_r^{(0)}\boldsymbol{y}_t = \Omega_r^{(0)}(\mathbf{Z}\boldsymbol{x}_t + \mathbf{a}) \\ &= \Omega_r^{(0)}(\mathbf{Z}\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{Z}\mathbf{I}_q^{(0)}\boldsymbol{x}_t + \mathbf{a}) \\ \boldsymbol{y}_t^+ &= \Omega_r^+\boldsymbol{y}_t = \Omega_r^+(\mathbf{Z}\boldsymbol{x}_t + \mathbf{a}) + \mathbf{v}_t^+ \\ &= \Omega_r^+(\mathbf{Z}\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{Z}\mathbf{I}_q^{(0)}\boldsymbol{x}_t + \mathbf{a}) + \mathbf{v}_t^+ \\ \mathbf{v}_t^+ &\sim \text{MVN}(0,\mathbf{R}^+)\end{aligned} \tag{161}$$

In order for this to be solvable using an EM algorithm with the Kalman filter, we require that no estimated $\mathbf{B}$ or $\mathbf{u}$ elements appear in the equation for $\boldsymbol{y}_t^{(0)}$. Since the $\boldsymbol{y}_t^{(0)}$ don't appear in the likelihood function ($\mathbf{R}^{(0)}=0$), $\boldsymbol{y}_t^{(0)}$ would not affect the the estimate for the parameters appearing in the $\boldsymbol{y}_t^{(0)}$ equation. This translates to the following constraints, $(\mathbf{I}\otimes\mathbf{Z}^+\mathbf{I}_q^{(0)})\mathbf{D}_B$ is all zeros and $(\mathbf{I}\otimes\mathbf{Z}^+\mathbf{I}_q^{(0)})\mathbf{D}_u$ is all zeros. Also notice that $\Omega_r^{(0)}\mathbf{Z}$ and $\Omega_r^{(0)}\mathbf{a}$ appear in the $\boldsymbol{y}^{(0)}$ equation and not in the $\boldsymbol{y}^+$ equation. This means that $\Omega_r^{(0)}\mathbf{Z}$ and $\Omega_r^{(0)}\mathbf{a}$ cannot be estimated but must be fixed terms.

In summary, the degenerate model becomes

$$\begin{aligned}
\boldsymbol{x}_t^{(0)} &= \mathbf{B}^{(0)}\boldsymbol{x}_{t-1} + \mathbf{u}^{(0)}\\
\boldsymbol{x}_t^{+} &= \mathbf{B}^{+}\boldsymbol{x}_{t-1} + \mathbf{u}^{+} + \mathbf{w}_t^{+}\\
\mathbf{w}_t^{+} &\sim \mathrm{MVN}(0,\mathbf{Q}^{+})\\
\boldsymbol{x}_0 &\sim \mathrm{MVN}(\xi,\Lambda)\\
\boldsymbol{y}_t^{(0)} &= \mathbf{Z}^{(0)}\mathbf{I}_q^{+}\boldsymbol{x}_t + \mathbf{Z}^{(0)}\mathbf{I}_q^{(0)}\boldsymbol{x}_t + \mathbf{a}^{(0)}\\
\boldsymbol{y}_t^{+} &= \mathbf{Z}^{+}\boldsymbol{x}_t + \mathbf{a}^{+} + \mathbf{v}_t^{+}\\
&= \mathbf{Z}^{+}\mathbf{I}_q^{+}\boldsymbol{x}_t + \mathbf{Z}^{+}\mathbf{I}_q^{(0)}\boldsymbol{x}_t + \mathbf{a}^{+} + \mathbf{v}_t^{+}\\
\mathbf{v}_t^{+} &\sim \mathrm{MVN}(0,\mathbf{R}^{+})
\end{aligned} \tag{162}$$

where $\mathbf{B}^{(0)} = \Omega_q^{(0)}\mathbf{B}$ and $\mathbf{B}^{+} = \Omega_q^{+}\mathbf{B}$ so that $\mathbf{B}^{(0)}$ are the rows of $\mathbf{B}$ corresponding to the diagonal of $\mathbf{Q} = 0$ and $\mathbf{B}^{+}$ are the rows of $\mathbf{B}$ corresponding to the diagonal of $\mathbf{Q} \neq 0$. The other parameters are similarly defined: $\mathbf{u}^{(0)} = \Omega_q^{(0)}\mathbf{u}$ and $\mathbf{u}^{+} = \Omega_q^{+}\mathbf{u}$, $\mathbf{Z}^{(0)} = \Omega_r^{(0)}\mathbf{Z}$ and $\mathbf{Z}^{+} = \Omega_r^{+}\mathbf{Z}$, and $\mathbf{a}^{(0)} = \Omega_r^{(0)}\mathbf{a}$ and $\mathbf{a}^{+} = \Omega_r^{+}\mathbf{a}$.

We want to write down the joint likelihood of $\boldsymbol{y}^{+} = \{\boldsymbol{y}_1^{+}, \boldsymbol{y}_2+, \boldsymbol{y}_3+, \dots\}$ and $\boldsymbol{x}^{+} = \{\boldsymbol{x}_1^{+}, \boldsymbol{x}_2^{+}, \boldsymbol{x}_3^{+}, \dots\}$. We can write the joint log-likelihood function for the + elements using equations 160 and 161 along with the likelihood function for a multivariate normal distribution.

$$\begin{aligned}
&\log \mathbf{L}(\boldsymbol{y}^{+}, \boldsymbol{x}^{+}; \Theta) =\\
&-\frac{1}{2}\sum_{1}^{T}(\boldsymbol{y}_t^{+} - \mathbf{Z}^{+}(\mathbf{I}_q^{+}\boldsymbol{x}_t + \mathbf{I}_q^{(0)}\boldsymbol{x}_t) - \mathbf{a}^{+})^\top(\mathbf{R}^{+})^{-1}\\
&\quad(\boldsymbol{y}_t^{+} - \mathbf{Z}^{+}(\mathbf{I}_q^{+}\boldsymbol{x}_t + \mathbf{I}_q^{(0)}\boldsymbol{x}_t) - \mathbf{a}^{+}) - \frac{T}{2}\log|\mathbf{R}^{+}|\\
&-\frac{1}{2}\sum_{1}^{T}(\boldsymbol{x}_t^{+} - \mathbf{B}^{+}\boldsymbol{x}_{t-1} - \mathbf{u}^{+})^\top(\mathbf{Q}^{+})^{-1}(\boldsymbol{x}_t^{+} - \mathbf{B}^{+}\boldsymbol{x}_{t-1} - \mathbf{u}^{+}) - \frac{T}{2}\log|\mathbf{Q}^{+}|\\
&-\frac{1}{2}(\boldsymbol{x}_0 - \xi)^\top\Lambda^{-1}(\boldsymbol{x}_0 - \xi) - \frac{1}{2}\log|\Lambda| - \frac{n}{2}\log 2\pi
\end{aligned} \tag{163}$$

n is the number of data points. If either $\mathbf{R}$ or $\mathbf{Q}$ are all zero, the line in the log-likelihood equation involving $\mathbf{R}^{+}$ or $\mathbf{Q}^{+}$ disappears. Notice that $\mathbf{a}^{(0)}$ and $\mathbf{Z}^{(0)}$ do not appear, which means that the rows of $\mathbf{a}$ and $\mathbf{Z}$ associated with deterministic $\boldsymbol{y}$ do not appear. Since these parameters do not appear in the likelihood (as written above), we cannot maximize the expected log-likelihood with respect to them. Notice also that $\mathbf{B}^{(0)}$ and $\mathbf{u}^{(0)}$ appear in the $\boldsymbol{y}$ part of the likelihood (in $\mathbf{I}_q^{(0)}\boldsymbol{x}_t$) while $\mathbf{B}^{+}$ and $\mathbf{u}^{+}$ appear in the $\boldsymbol{x}$ part.

If $\boldsymbol{x}_0$ is treated as fixed ($\Lambda = 0$), then the likelihood takes a slightly different form using equation (162)

$$
\begin{aligned}
&\log \mathbf{L}(\boldsymbol{y}^+,\boldsymbol{x}^+;\Theta) = \\
&-\frac{1}{2}\sum_1^T(\boldsymbol{y}_t^+ - \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{I}_q^{(0)}\boldsymbol{x}_t) - \mathbf{a}^+)^\top(\mathbf{R}^+)^{-1} \\
&\quad(\boldsymbol{y}_t^+ - (\mathbf{Z}^+\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{I}_q^{(0)}\boldsymbol{x}_t) - \mathbf{a}^+) - \frac{T}{2}\log|\mathbf{R}^+| \\
&-\frac{1}{2}\sum_1^T(\boldsymbol{x}_t^+ - \mathbf{B}^+\boldsymbol{x}_{t-1} - \mathbf{u}^+)^\top(\mathbf{Q}^+)^{-1}(\boldsymbol{x}_t^+ - \mathbf{B}^+\boldsymbol{x}_{t-1} - \mathbf{u}^+) \\
&-\frac{T}{2}\log|\mathbf{Q}^+| - \frac{n}{2}\log 2\pi \\
&\quad\text{where } \boldsymbol{x}_0 \equiv \xi
\end{aligned}
\tag{164}
$$

If the initial condition parameters refer to $\boldsymbol{x}_1$ instead of $\boldsymbol{x}_0$, then the $\boldsymbol{x}$ summation in equations 164 and 163 starts at 2 instead of 1.

6.2.2 Summary of requirements

Below are discussed the update equations for the different parameters. Here I summarize the constraints that are scattered throughout these subsections.

- $(\mathbf{I}\otimes\mathbf{Z}^+\mathbf{I}_q^{(0)})\mathbf{D}_B$ is all zeros; if there is a y with $\mathbf{R}=0$ and it is linked (through $\mathbf{Z}$) to an x with $\mathbf{Q}=0$, then the corresponding $\mathbf{B}$ elements are fixed instead of estimated.
- $(\mathbf{I}\otimes\mathbf{Z}^+\mathbf{I}_q^{(0)})\mathbf{D}_u$ is all zeros; if there is a y with $\mathbf{R}=0$ and it is linked (through $\mathbf{Z}$) to an x with $\mathbf{Q}=0$, then the corresponding $\mathbf{u}$ elements are fixed instead of estimated.
- $(\mathbf{I}\otimes\Omega_r^{(0)})\mathbf{D}_z$ is all zeros; if y has no observation error, then the corresponding $\mathbf{Z}$ rows are fixed values.
- $(\mathbf{I}\otimes\Omega_r^{(0)})\mathbf{D}_a$ is all zeros; if y has no observation error, then the corresponding $\mathbf{a}$ rows are fixed values.
- $\Omega_r^+\mathbf{Z}\mathbf{I}_q^{(0)}$ has no rows that are all zeros; this is a sufficient contraint but it's a bit too stringent. The constraint is that the $\mathbf{B}^{(0)}$ elements appear in the $\boldsymbol{y}^+$ part of the likelihood and you need to make sure that all the elements appear (aren't zero-ed out by zero rows in $\Omega_r^+\mathbf{Z}\mathbf{I}_q^{(0)}$). This requirement means that there are no deterministic processes observed with no errors; if you have deterministic processes that are observed with no errors, then you should be able to rewrite the model to remove that redundant y by having $n<m$.
- $\mathbf{B}^{(0)}$ is fixed. While it could be estimated potentially, the derivation here assumes it is not.

- If part or all of $\mathbf{u}^{(0)}$ is estimated, then the estimated $\mathbf{u}^{(0)}$ elements must unconnected to the stochastic part of the $\boldsymbol{x}$ model; this allows us to use the matrix geometric series to rewrite $\boldsymbol{x}^{(0)}$ in terms of $\mathbf{u}^{(0)}$, $\xi^{(0)}$ and $\mathbf{B}^{(0)}$ in the $\mathbf{u}$ update equation. $\mathbf{B}$ must be block diagonal with $\Omega_q^{(0)}\mathbf{B}(\Omega_q^{(0)})^\top$ and $\Omega_q^{+}\mathbf{B}(\Omega_q^{+})^\top$ in separate blocks or . This means that deterministic state processes are not linked to the stochastic state processes through $\mathbf{B}$. Also the absolute value of all the eigenvalues of $\Omega_q^{(0)}\mathbf{B}(\Omega_q^{(0)})^\top$ must be less than 1.
- If $\mathbf{u}^(0)$ is fixed, then $\mathbf{B}$ need not be block diagonal. The only requirement is that $\mathbf{B}^{(0)}$ is fixed.

The dimension of the identity matrices in the above constraints is implicit.

6.2.3 $\mathbf{Z}^{+}$ and $\mathbf{a}^{+}$ update equations for partially deterministic models

The $\mathbf{a}$ and $\mathbf{Z}$ update equations involve both $\widetilde{\mathbf{y}}_t$ and the inverse of $\mathbf{R}$ and thus must be modified allow zeros on the diagonal of $\mathbf{R}$.

Because we require that $\mathbf{Z}^{(0)}$ and $\mathbf{a}^{(0)}$ are fixed, we can rewrite the $\mathbf{Z}$ update equation in the case where there are zeros on the diagonal of $\mathbf{R}$ as the constrained update equation for $\mathbf{Z}$ (equation 111) with $\mathbf{R}^{-1}$ replaced with $\mathbf{R}^{*}$:

$$\begin{aligned}\boldsymbol{\zeta}_{j+1} = \bigg(\sum_{t=1}^{T}(\mathbf{D}_z^\top(\widetilde{\mathbf{P}}_t\otimes\mathbf{R}^*)\mathbf{D}_z)\bigg)^{-1}\mathbf{D}_z^\top\times \\ \sum_{t=1}^{T}\big(\mathrm{vec}(\mathbf{R}^*(\widetilde{\mathbf{yx}}_t-\mathbf{a}\widetilde{\mathbf{x}}_t^\top))-(\widetilde{\mathbf{P}}_t\otimes\mathbf{R}^*)\mathbf{f}_z\big)\end{aligned} \tag{165}$$

where $\mathbf{R}^* = (\Omega_r^{+})^\top(\mathbf{R}^{+})^{-1}\Omega_r^{+}$. Combining $\boldsymbol{\zeta}_{j+1}$ with $\mathbf{Z}_{\text{fixed}}$, we arrive at the vec of the updated $\mathbf{Z}$ matrix:

$$\mathrm{vec}(\mathbf{Z}_{j+1}) = \mathbf{f}_z + \mathbf{D}_z\boldsymbol{\zeta}_{j+1} \tag{166}$$

Because the $\mathbf{Z}^{(0)}$ elements are fixed, $\mathbf{D}_z^\top(\widetilde{\mathbf{P}}_t\otimes\mathbf{R}^*)\mathbf{D}_z$ is invertible. As usual, $\mathbf{Z}$ elements must be fixed in such a way that the model has one solution.

Similarly, the derivation for the constrained $\mathbf{a}$ update equation also reduces to the constrained $\mathbf{a}$ equation (equation 88) with $\mathbf{R}^{-1}$ replaced with $\mathbf{R}^{*}$:

$$\boldsymbol{\alpha}_{j+1} = \frac{1}{T}\big(\mathbf{D}_a^\top\mathbf{R}^*\mathbf{D}_a\big)^{-1}\mathbf{D}_a^\top\mathbf{R}^*\sum_{t=1}^{T}\big(\widetilde{\mathbf{y}}_t-\mathbf{Z}\widetilde{\mathbf{x}}_t-\mathbf{f}_a\big) \tag{167}$$

The new $\mathbf{a}$ parameter is then

$$\mathbf{a}_{j+1} = \mathbf{f}_a + \mathbf{D}_a\boldsymbol{\alpha}_{j+1}, \tag{168}$$

The $\mathbf{a}^{(0)}$ elements are fixed which means that $\mathbf{D}_a^\top\mathbf{R}^*\mathbf{D}_a$ is invertible. For example, if $\mathbf{R}$ is all zeros and $\mathbf{Z}$ is a column vector, then all the $\mathbf{a}$ elements must be fixed.

6.2.4 Systems with fully deterministic $\boldsymbol{x}$ rows

Our process equation is $\boldsymbol{x}_t = \mathbf{B}\boldsymbol{x}_{t-1} + \mathbf{u}$, with the $\mathbf{w}_t$ term left off. Each row i in $\mathbf{u}$ is an individual u_i parameter although we work with $\mathbf{u}$ as a vector. Each u_i is associated with row i in $\boldsymbol{x}_t$ (left side). When we do the partial differentiation step in deriving the EM update equation for $\mathbf{u}$ (or ξ if $\Lambda = 0$), we will need to take a partial derivative while holding $\boldsymbol{x}_t$ (which includes $\boldsymbol{x}_{t-1}$) constant. If any of our rows in $\boldsymbol{x}_t$ are fully deteministic, meaning no process variance for that row and NOT connected through $\mathbf{B}$ to any of the stochastic rows, then we cannot hold that row of $\boldsymbol{x}$ constant while changing the corresponding row of $\mathbf{u}$ (or ξ if $\Lambda = 0$). If a row of $\boldsymbol{x}$ is fully deterministic, then that x_i must change when u_i is changed (or ξ_i if $\Lambda = 0$). Thus we simply cannot do the partial differentiation step required in the EM update equation derivation.

So we need to identify the fully deterministic $\boldsymbol{x}$ and treat them differently in our update equation. x_i is directly stochastic when the corresponding $\mathbf{Q}$ diagonal term is non-zero. I will denote this group as the $\mathbf{Q} \neq 0$ group. Secondly, x_i could be indirectly stochastic by being connected to the $\mathbf{Q} \neq 0$ group through $\mathbf{B}$. If we replace all non-zero elements in $\mathbf{B}$ with 1, then we have an adjacency matrix, let's call it $\mathbf{M}$, for our system of x's. Then finding out whether x_i is fully deterministic is a matter of determining if there exists a path in $\mathbf{M}$ from x_i in the $\mathbf{Q} = 0$ group to the x_i in the $\mathbf{Q} \neq 0$ group. If no path exists, then x_i is fully deterministic. Note if x_i is fully deterministic but we are not trying to estimate u_i or ξ_i or $\mathbf{B}_{i,\cdot}$, then it does not matter since we won't be taking the partial derivative with respect to u_i.

Denote the i for the fully deterministic $\boldsymbol{x}$ rows as d. $\mathbf{B}$ can be thought of has having two types of rows: stochastic either directly on indirectly and fully deterministic. By definition, the $\mathbf{B}$ can be rearranged to look something like so where s are stochastic because $\mathbf{Q} \neq 0$, is are indirectly stochastic because $\mathbf{Q} = 0$ but they are linked to the s rows through $\mathbf{B}$, and d are fully deterministic rows because $\mathbf{Q} = 0$ and they are not linked to the s rows through $\mathbf{B}$:

$$\begin{bmatrix} s & s & s & s & s \\ s & s & s & s & s \\ is & is & is & is & is \\ 0 & 0 & d & d & d \\ 0 & 0 & d & d & d \end{bmatrix} \tag{169}$$

The s's, is's and d's are not all equal; I am just showing the blocks. The 0s in the fully deterministic rows are what is causing these to be fully deterministic. This is the $\mathbf{Q} = 0 \bullet \Lambda = 0$ group that is unconnected to the $\mathbf{Q} \neq 0$ group through any path through $\mathbf{B}$.

How do you determine the d, or deterministic, set of $\boldsymbol{x}$ rows? Since my $\mathbf{B}$ matrices are small, I use a very inefficient strategy in my code. I define the $\mathbf{M}$ adjacency matrix by replacing all non-zero $\mathbf{B}$ values with 1. Then I raise $\mathbf{M}$ to the m-th power to find all the connections of length m or smaller. I subset out the $\mathbf{M}^m$ rows associated with $\mathbf{Q} = 0$. Within that subset, those rows where only 0s appear in the $\mathbf{Q} \neq 0$ columns are the fully deterministic $\boldsymbol{x}$ rows. This is inefficient because taking the m-th power of $\mathbf{M}$ is slow as m gets large. There are much faster algorithm for finding paths, but this test is only done once at the start of the EM algorithm.

6.2.5 $\mathbf{u}$ update equation for systems with fully deterministic $\boldsymbol{x}$ rows

To derive the update equation for $\mathbf{u}$, we need to take the partial derivative of Ψ^+ holding everything constant except $\mathbf{u}$, which includes both $\mathbf{u}^d$ and $\mathbf{u}^s$ (d denotes the fully deterministic and s denotes stochastic, directly or indirectly). The state processes in $\boldsymbol{x}^d$ are fully deterministic, therefore we cannot hold $\boldsymbol{x}_t^d$ constant while changing $\mathbf{u}^d$. If we change $\mathbf{u}^d$, then $\boldsymbol{x}_t^d$ must change because it is fully deterministic. This is in contrast to $\mathbf{u}^s$ which can be changed while holding $\boldsymbol{x}^s$ constant, because $\boldsymbol{x}^s$ is stochastic (perhaps indirectly through $\mathbf{B}$) and all values are possible for a given $\mathbf{u}^s$ (and the $\boldsymbol{x}^d$ are unconnected to $\boldsymbol{x}^s$ so also do not change if $\mathbf{u}^s$ is changed). Thus we need to replace $\boldsymbol{x}_t^d$ (it is inside $\mathbf{I}_q^{(0)}\boldsymbol{x}_t$) appearing in the likelihood with an equation that does not involve $\boldsymbol{x}_t^d$.

By definition, all the $\mathbf{B}$ elements in the s columns of $\mathbf{B}^d$ are 0 (see equation 169). If the absolute value of all the eigenvalues of $\mathbf{B}^{d,d}$ are less than 1 ($\mathbf{B}^{d,d}$ is the block of d's in equation 169), we can rewrite the equation for $\boldsymbol{x}^d$ as follows using the matrix geometric series:

$$\begin{aligned}\boldsymbol{x}_t^d =&(\mathbf{B}^{d,d})^t\boldsymbol{x}_0^d + \sum_{i=0}^{t-1}(\mathbf{B}^{d,d})^i\mathbf{u}^d = \\ &(\mathbf{B}^{d,d})^t\boldsymbol{x}_0^d + (\mathbf{I}-\mathbf{B}^{d,d})^{-1}(\mathbf{I}-(\mathbf{B}^{d,d})^t)\mathbf{u}^d, \quad \text{if } \mathbf{B}^{d,d} \neq \mathbf{I} \\ &\boldsymbol{x}_0^d + \mathbf{u}^d t, \quad \text{if } \mathbf{B}^{d,d} = \mathbf{I}\end{aligned} \tag{170}$$

where $\mathbf{B}^{d,d}$ is the block of d's in equation 169.

Then to obtain the $\mathbf{u}$ update equation, we will replace the $\mathbf{I}_q^{(0)}\boldsymbol{x}_t$ term appearing in the likelihood (equation 163) with $\mathbf{I}_q^d\boldsymbol{x}_t + \mathbf{I}_q^{is}\boldsymbol{x}_t$ where $\mathbf{I}_q^d$ and $\mathbf{I}_q^{is}$ are defined in the standard way: they are diagonal matrices where all elements are 0 except the diagonals corresponding to the d (or is) rows are set to 1. $\mathbf{I}_q^{(0)} = \mathbf{I}_q^d + \mathbf{I}_q^{is}$ since for the d and is rows are associated with $\mathbf{Q}$ diagonals equal to 0. We are only concerned about the $\mathbf{I}_q^d\boldsymbol{x}_t$ because these are the $\boldsymbol{x}$ elements that cannot be held fixed when $\mathbf{u}^d$ is changed.

We rewrite this with the matrix geometric series:

$$\begin{aligned}\mathbf{I}_q^d\boldsymbol{x}_t &= (\mathbf{B}^\bullet)^t\boldsymbol{x}_0 + \mathbf{I}_q^d(\mathbf{I}_m - \mathbf{B}^\bullet)^{-1}(\mathbf{I}_m - (\mathbf{B}^\bullet)^t)\mathbf{I}_q^d\mathbf{u} \\ &= \mathbf{B}^\diamondsuit\boldsymbol{x}_0 + \mathbf{B}^\sharp\mathbf{u} \\ &\text{where } \mathbf{B}^\bullet = \mathbf{I}_q^d\mathbf{B}\mathbf{I}_q^d \\ &\text{where } \mathbf{B}^\diamondsuit = (\mathbf{B}^\bullet)^t \\ &\text{and } \mathbf{B}^\sharp = \mathbf{I}_q^d(\mathbf{I}_m - \mathbf{B}^\bullet)^{-1}(\mathbf{I}_m - \mathbf{B}^\diamondsuit)\mathbf{I}_q^d\end{aligned} \tag{171}$$

The Equation 171 has been written slightly differently so that we are working with a $\mathbf{B}$ matrix with the stochastic row/columns zeroed out ($\mathbf{I}_q^d\mathbf{B}\mathbf{I}_q^d$) and the $\boldsymbol{x}_t$ vector with the stochastic rows zeroed out ($\mathbf{I}_q^d\boldsymbol{x}_t$). If any block of $\mathbf{B}^\bullet$ is an identity matrix, then we will need to replace the corresponding block in $(\mathbf{B}^\diamondsuit\boldsymbol{x}_0 + \mathbf{B}^\sharp)$ with $t\mathbf{I}$, a diagonal block with t on the diagonal.

We will replace $\mathbf{I}_q^{is}\boldsymbol{x}_t$ with $\mathbf{I}_q^{is}(\mathbf{B}\boldsymbol{x}_{t-1} + \mathbf{u})$. We cannot do this with the deterministic rows in $\boldsymbol{x}$ because when we try to do the partial differentiation, we would not be able to hold $\boldsymbol{x}_{t-1}^{d}$ constant since it is fully deterministic. But since $\boldsymbol{x}_{t-1}^{is}$ is stochastic (indirectly through $\mathbf{B}$) we can do the partial differentiation step.

Thus Ψ^{+} becomes

$$\begin{aligned}
&\Psi^{+} = \mathrm{E}[\log \mathbf{L}(\boldsymbol{Y}^{+},\boldsymbol{X}^{+};\Theta)] = \\
&\mathrm{E}[-\frac{1}{2}\sum_{1}^{T}(\boldsymbol{Y}_t^{+} - \mathbf{Z}^{+}(\mathbf{I}_q^{+}\boldsymbol{X}_t + \mathbf{I}_q^{is}(\mathbf{B}\boldsymbol{X}_{t-1} + \mathbf{u}) + \mathbf{I}_q^{d}(\mathbf{B}^{\diamondsuit}\boldsymbol{X}_0 + \mathbf{B}^{\sharp}\mathbf{u})) - \mathbf{a}^{+})^{\top}(\mathbf{R}^{+})^{-1} \\
&\quad (\boldsymbol{Y}_t^{+} - \mathbf{Z}^{+}(\mathbf{I}_q^{+}\boldsymbol{X}_t + \mathbf{I}_q^{is}(\mathbf{B}\boldsymbol{X}_{t-1} + \mathbf{u}) + \mathbf{I}_q^{d}(\mathbf{B}^{\diamondsuit}\boldsymbol{x}_0 + \mathbf{B}^{\sharp}\mathbf{u})) - \mathbf{a}^{+}) - \frac{T}{2}\log|\mathbf{R}^{+}| \\
&-\frac{1}{2}\sum_{1}^{T}(\boldsymbol{X}_t^{+} - \mathbf{B}^{+}\boldsymbol{X}_{t-1} - \mathbf{u}^{+})^{\top}(\mathbf{Q}^{+})^{-1}(\boldsymbol{X}_t^{+} - \mathbf{B}^{+}\boldsymbol{X}_{t-1} - \mathbf{u}^{+}) \\
&-\frac{T}{2}\log|\mathbf{Q}^{+}| - \frac{1}{2}(\boldsymbol{X}_0 - \xi)^{\top}\Lambda^{-1}(\boldsymbol{X}_0 - \xi) - \frac{1}{2}\log|\Lambda| - \frac{n}{2}\log 2\pi
\end{aligned} \tag{172}$$

The $\mathbf{u}^{(0)}$ parameter appears in the $\boldsymbol{Y}$ part of the likelihood (as $(\mathbf{I}_q^{is}+\mathbf{I}_q^{d})u$) and $\mathbf{u}^{+}$ appears in the $\boldsymbol{x}$ part. However, because $\mathbf{u}$ can have shared elements, it is possible that a $\mathbf{u}$ element is shared across $\mathbf{u}^{(0)}$ and $\mathbf{u}^{+}$. We write $\mathbf{u}$ as $\mathbf{f}_u + \mathbf{D}_u\boldsymbol{\upsilon}$, put that in equation (172), and differentiate with respect to $\boldsymbol{\upsilon}$ rather than $\mathbf{u}^{(0)}$ or $\mathbf{u}^{+}$.

The derivation steps are similar to those for the general update equation (analogous to equation 86). Take the derivative of Ψ^{+} (equation 172) with respect to $\boldsymbol{\upsilon}$. After taking the derivative with respect to $\boldsymbol{\upsilon}$, we get:

$$\begin{aligned}
&\mathbf{D}_u^{\top}(\mathbf{R}^{\sharp} + T\mathbf{Q}^{*})\mathbf{D}_u\boldsymbol{\upsilon} = \\
&\quad \mathbf{D}_u^{\top}\mathbf{I}_q^{(0)}\sum_{t=1}^{T}(\mathbf{B}^{\sharp} + \mathbf{I}_q^{is})^{\top}\mathbf{Z}^{\top}\mathbf{R}^{*}\big(\widetilde{\mathbf{y}}_t - \mathbf{Z}\mathbf{I}_q^{+}\widetilde{\mathbf{x}}_t - \mathbf{Z}\mathbf{I}_q^{is}\mathbf{B}\widetilde{\mathbf{x}}_{t-1} - \mathbf{Z}\mathbf{I}_q^{is}\mathbf{f}_u - \mathbf{Z}\mathbf{I}_q^{d}(\mathbf{B}^{\diamondsuit}\widetilde{\mathbf{x}}_0 + \mathbf{B}^{\sharp}\mathbf{f}_u) - \mathbf{a}\big) \\
&\quad + \mathbf{D}_u^{\top}\mathbf{I}_q^{+}\mathbf{Q}^{*}\sum_{t=1}^{T}\big(\widetilde{\mathbf{x}}_t - \mathbf{B}\widetilde{\mathbf{x}}_{t-1} - \mathbf{f}_u\big) \\
&\text{where } \mathbf{R}^{*} = (\Omega_r^{+})^{\top}(\mathbf{R}^{+})^{-1}\Omega_r^{+} \\
&\text{and } \mathbf{R}^{\sharp} = \sum_{t=1}^{T}(\mathbf{B}^{\sharp} + \mathbf{I}_q^{is})^{\top}\mathbf{Z}^{\top}\mathbf{R}^{*}\mathbf{Z}(\mathbf{B}^{\sharp} + \mathbf{I}_q^{is}) \\
&\text{and } \mathbf{Q}^{*} = (\Omega_q^{+})^{\top}(\mathbf{Q}^{+})^{-1}\Omega_q^{+}
\end{aligned} \tag{173}$$

$\mathbf{R}^{\sharp}$ will have 0s where $\mathbf{Q} \neq 0$ and $\mathbf{Q}^{*}$ will have 0s where $\mathbf{Q} = 0$. $\mathbf{B}^{\sharp}$ is wrapped in $\mathbf{I}_q^{d}$ which is why that does not appear in front of it in the equation. Note that $\mathbf{R}^{\sharp} + \mathbf{Q}^{*}$ does not have any zero rows or columns since we require that any state process with zero variance is observed with errors (no zero rows in $\Omega_r^{+}\mathbf{Z}\mathbf{I}_q^{(0)}$). This means that corresponding $\mathbf{Q}^{(0)}$ rows/columns of $(\mathbf{B}^{\sharp})^{\top}\mathbf{Z}^{\top}\mathbf{R}^{*}\mathbf{Z}\mathbf{B}^{\sharp}$ will be non-zero. Note that the $\mathbf{B}^{(00)}$ part of $\mathbf{B}^{\sharp}$ will never have all zero rows/columns by virtue of its definition. Also note that because $\mathbf{Q}^{*} = \mathbf{I}_q^{+}\mathbf{Q}^{*}\mathbf{I}_q^{+}$ by definition, $\mathbf{R}^{\sharp}$ is contributing to the $u's$ associated with $\mathbf{Q} = 0$ and $\mathbf{Q}^{*}$ contributes to the $u's$ associated with $\mathbf{Q} \neq 0$.

Thus, the updated $\boldsymbol{\upsilon}$ is

$$\begin{aligned}\boldsymbol{\upsilon}_{j+1} = &\big(\mathbf{D}_u^\top(\mathbf{R}^\sharp + T\mathbf{Q}^*)\mathbf{D}_u\big)^{-1}\mathbf{D}_u^\top \times \\ &\bigg(\sum_{t=1}^{T}(\mathbf{B}^\sharp + \mathbf{I}_q^{is})^\top\mathbf{Z}^\top\mathbf{R}^*\big(\widetilde{\mathbf{y}}_t - \mathbf{Z}\mathbf{I}_q^+\widetilde{\mathbf{x}}_t - \mathbf{Z}\mathbf{I}_q^{is}\mathbf{B}\widetilde{\mathbf{x}}_{t-1} - \mathbf{Z}\mathbf{I}_q^{is}\mathbf{f}_u - \mathbf{Z}\mathbf{B}^\diamond\widetilde{\mathbf{x}}_0 - \mathbf{Z}\mathbf{B}^\sharp\mathbf{f}_u - \mathbf{a}\big) \\ &+ \mathbf{I}_q^+\mathbf{Q}^*\sum_{t=1}^{T}\big(\widetilde{\mathbf{x}}_t - \mathbf{B}\widetilde{\mathbf{x}}_{t-1} - \mathbf{f}_u\big)\bigg)\end{aligned} \tag{174}$$

and

$$\mathbf{u}_{j+1} = \mathbf{f}_u + \mathbf{D}_u\boldsymbol{\upsilon}_{j+1}, \tag{175}$$

where $\mathbf{B}^\diamond$ and $\mathbf{B}^\sharp$ are defined in equation (171) and $\mathbf{R}^\sharp$, $\mathbf{R}^*$ and $\mathbf{Q}^*$ are defined in equation (173). If $\boldsymbol{x}_0$ is treated as fixed, $\widetilde{\mathbf{x}}_0$ is replaced with ξ, otherwise it has its usual definition ($\mathrm{E}[\boldsymbol{X}_0|\boldsymbol{y}(1),\Theta_j]$).

Conceptually, I think the approach described here is the same as the approach presented in section 4.2.5 of (Harvey, 1989), but it is more general because it deals with the case where some $\mathbf{u}$ elements are shared (linear functions of some set of shared values), possibly across deterministic and stochastic elements. Also, I present it here within the context of the EM algorithm, so solving for the maximum-likelihood $\mathbf{u}$ appears in the context of maximizing Ψ^+ with respect to $\mathbf{u}$ for the update equation at iteration $j+1$.

6.2.6 $\mathbf{u}^+$ update equation when $\mathbf{u}^{(0)}$ is not estimated

When $\mathbf{u}^{(0)}$ is not estimated (since it is at some user defined value), we do not need to take the partial derivative with respect to $\mathbf{u}^{(0)}$. Thus the likelihood (equation 163) is unchanged. Since $\mathbf{u}^+$ only appears in the $\boldsymbol{x}$ part of the likelihood, the update equation for $\mathbf{u}$ is relatively unchanged: Thus,

$$\boldsymbol{\upsilon}_{j+1} = \frac{1}{T}\big(\mathbf{D}_u^\top\mathbf{Q}^*\mathbf{D}_u\big)^{-1}\mathbf{D}_u^\top\mathbf{Q}^*\sum_{t=1}^{T}\big(\widetilde{\mathbf{x}}_t - \mathbf{B}\widetilde{\mathbf{x}}_{t-1} - \mathbf{f}_u\big) \tag{176}$$

and

$$\mathbf{u}_{j+1} = \mathbf{f}_u + \mathbf{D}_u\boldsymbol{\upsilon}_{j+1}, \tag{177}$$

The difference is that $\mathbf{Q}^*$ appears in the equation instead of $\mathbf{Q}^{-1}$ to deal with the 0s on the diagonal of $\mathbf{Q}$ when taking the inverse. Equation 175 reduces to this since $\mathbf{D}_u^\top\mathbf{I}_q^{(0)}$ will be all zeros so the $\mathbf{R}$ part of the update equation drops out.

6.2.7 ξ^+ update equation when $\xi^{(0)}$ is not estimated

When $\xi^{(0)}$ is not estimated (because you fixed it as some value), we do not need to take the partial derivative with respect to $\xi^{(0)}$ since we will not be estimating it. Thus the likelihood (equation 163) is unchanged and the update equation for ξ is relatively unchanged (see section on the unconstrained ξ update equations. The one difference is that $\mathbf{Q}^*$ will appear in the update equation to deal with the 0s on the diagonal of $\mathbf{Q}$ when taking the inverse.

6.2.8 ξ update equation when $\Lambda \neq 0$

If $\Lambda^{(0)} \neq 0$ then the update equation for ξ does not change since we can take the partial derivative of Ψ^+ while holding $\boldsymbol{X}_0$ constant.

6.2.9 $\xi^{(0)}$ update equation when $\Lambda = 0$ and $\boldsymbol{x}_0 \equiv \xi$

Define $\xi^{(0)}$ as the ξ rows corresponding to $\mathbf{Q}$ diagonal values equal to 0. The expected log likelihood function for this case, when $\Lambda = 0$ is written:

$$\begin{aligned}
&\Psi^+ = \mathrm{E}[\log \mathbf{L}(\boldsymbol{Y}^+,\boldsymbol{X}^+;\Theta)] = \\
&\mathrm{E}[-\frac{1}{2} - \sum_1^T (\boldsymbol{Y}_t^+ - \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{X}_t^+ + \mathbf{I}_q^{(0)}\boldsymbol{X}_t^{(0)}) - \mathbf{a}^+)^\top (\mathbf{R}^+)^{-1} (\boldsymbol{Y}_t^+ - \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{X}_t^+ + \mathbf{I}_q^{(0)}\boldsymbol{X}_t^{(0)}) - \mathbf{a}^+) - \sum_1^T \frac{1}{2}\log|\mathbf{R}^+| \\
&\quad - \sum_1^T \frac{1}{2}(\boldsymbol{X}_t^+ - \mathbf{B}^+\boldsymbol{X}_{t-1} - \mathbf{u}^+)^\top (\mathbf{Q}^+)^{-1} (\boldsymbol{X}_t^+ - \mathbf{B}^+\boldsymbol{X}_{t-1} - \mathbf{u}^+) - \sum_1^T \frac{1}{2}\log|\mathbf{Q}^+| - \frac{n}{2}\log 2\pi] \\
&\boldsymbol{x}_0 \equiv \xi
\end{aligned} \tag{178}$$

When $\Lambda^{(0)} = 0$, we run into the same troubles as for $\mathbf{u}^{(0)}$. We need to take the partial derivative of Ψ^+ holding everything but ξ constant. But if some of the $\boldsymbol{x}$ are fully deterministic, they will automatically change with ξ^d since they are a deterministic function of ξ^d; they won't be affected by the other ξ rows since by definition fully deterministic $\boldsymbol{x}$ are not connected, via $\mathbf{B}$, to any other rows except those in the fully deterministic group. If some of the $\boldsymbol{x}$ are indirectly stochastic, $\boldsymbol{x}_1^{is}$ will automatically change with ξ since $\boldsymbol{x}_1^{is} = \mathbf{B}\xi + \mathbf{u}$ (no $\mathbf{w}_t$).

We use the same trick as for $\mathbf{u}^{(0)}$:

$$\begin{aligned}
&\Psi^+ = \mathrm{E}[\log \mathbf{L}(\boldsymbol{Y}^+,\boldsymbol{X}^+;\Theta)] = \\
&\mathrm{E}[-(\boldsymbol{Y}_1^+ - \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{X}_1^+ + \mathbf{I}_q^{is}(\mathbf{B}\xi + \mathbf{u}) + \mathbf{I}_q^d(\mathbf{B}_1^\diamondsuit\xi + \mathbf{B}_1^\sharp\mathbf{u})) - \mathbf{a}^+)^\top \\
&(\mathbf{R}^+)^{-1}(\boldsymbol{Y}_1^+ - \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{X}_1 + \mathbf{I}_q^{is}(\mathbf{B}\xi + \mathbf{u}) + \mathbf{I}_q^d(\mathbf{B}_1^\diamondsuit\xi + \mathbf{B}_1^\sharp\mathbf{u})) - \mathbf{a}^+) \\
&-\sum_2^T (\boldsymbol{Y}_t^+ - \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{X}_t + \mathbf{I}_q^{is}\boldsymbol{X}_t + \mathbf{I}_q^d(\mathbf{B}_t^\diamondsuit\xi + \mathbf{B}_t^\sharp\mathbf{u})) - \mathbf{a}^+)^\top (\mathbf{R}^+)^{-1} (\boldsymbol{Y}_t^+ - \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{X}_t + \mathbf{I}_q^{is}\boldsymbol{X}_t + \mathbf{I}_q^d(\mathbf{B}_t^\diamondsuit\xi + \mathbf{B}_t^\sharp\mathbf{u})) - \mathbf{a}^+) - \sum_1^T \frac{1}{2}\mathrm{l} \\
&\quad - \sum_1^T \frac{1}{2}(\boldsymbol{X}_t^+ - \mathbf{B}^+\boldsymbol{X}_{t-1} - \mathbf{u}^+)^\top (\mathbf{Q}^+)^{-1} (\boldsymbol{X}_t^+ - \mathbf{B}^+\boldsymbol{X}_{t-1} - \mathbf{u}^+) - \sum_1^T \frac{1}{2}\log|\mathbf{Q}^+| - \frac{n}{2}\log 2\pi] \\
&\boldsymbol{x}_0 \equiv \xi
\end{aligned} \tag{179}$$

We take the derivative of Ψ^+ (equation 179) with respect to $\mathbf{p}$ where $\xi = \mathbf{f}_\xi + \mathbf{D}_\xi\mathbf{p}$.

The constrained $\mathbf{p}$ update equation when $\mathbf{Q}$ has zeros on the diagonal is then

$$\begin{aligned}
&\mathbf{D}_\xi^\top(\mathbf{R}^\Diamond+\mathbf{B}^\top\mathbf{Q}^*\mathbf{B})\mathbf{D}_\xi\mathbf{p}=\\
&\quad\mathbf{D}_\xi^\top\bigg(\mathbf{I}_q^{(0)}\sum_{t=1}^T\mathbf{B}^\Diamond\mathbf{Z}^\top\mathbf{R}^*\big(\widetilde{\mathbf{y}}_t-\mathbf{Z}\mathbf{I}_q^+\widetilde{\mathbf{x}}_t-\mathbf{Z}\mathbf{B}^\sharp\mathbf{u}-\mathbf{Z}\mathbf{B}^\Diamond\mathbf{f}_\xi-\mathbf{a}\big)\\
&\quad+\mathbf{I}_q^+\mathbf{B}^\top\mathbf{Q}^*\big(\widetilde{\mathbf{x}}_1-\mathbf{B}\mathbf{f}_\xi-\mathbf{u}\big)\bigg)\\
&\text{where }\mathbf{R}^\Diamond=\sum_{t=1}^T\mathbf{B}^\Diamond\mathbf{Z}^\top\mathbf{R}^*\mathbf{Z}\mathbf{B}^\Diamond
\end{aligned}\tag{180}$$

The matrices $\mathbf{B}^\Diamond$ and $\mathbf{B}^\sharp$ are defined in equation (171), and $\mathbf{R}^*$ and $\mathbf{Q}^*$ are defined in equation (173). The absolute value of all the eigenvalues of $\mathbf{B}^{(00)}$ are constrained to be less than or equal to 1.

Thus, the updated $\mathbf{p}$ is

$$\begin{aligned}
\mathbf{p}_{j+1}=&\big(\mathbf{D}_\xi^\top(\mathbf{R}^\Diamond+\mathbf{B}^\top\mathbf{Q}^*\mathbf{B})\mathbf{D}_\xi\big)^{-1}\mathbf{D}_\xi^\top\times\\
&\bigg(\mathbf{I}_q^{(0)}\sum_{t=1}^T\mathbf{B}^\Diamond\mathbf{Z}^\top\mathbf{R}^*\big(\widetilde{\mathbf{y}}_t-\mathbf{Z}\mathbf{I}_q^+\widetilde{\mathbf{x}}_t-\mathbf{Z}\mathbf{B}^\sharp\mathbf{u}-\mathbf{Z}\mathbf{B}^\Diamond\mathbf{f}_\xi-\mathbf{a}\big)\\
&\quad+\mathbf{I}_q^+\mathbf{B}^\top\mathbf{Q}^*\big(\widetilde{\mathbf{x}}_1-\mathbf{B}\mathbf{f}_\xi-\mathbf{u}\big)\bigg)
\end{aligned}\tag{181}$$

and

$$\xi_{j+1}=\mathbf{f}_\xi+\mathbf{D}_\xi\mathbf{p}_{j+1},\tag{182}$$

6.2.10 ξ^+ update equation when $\xi^{(0)}$ is fixed

When $\xi^{(0)}$ is fixed, we do not need to take the partial derivative with respect to $\xi^{(0)}$ since it is now fixed. Thus the likelihood (equation 163) is unchanged and the update equation for ξ is relatively unchanged (see section on the unconstrained ξ update equations.

For example if ξ is treated as an unknown parameters and $\Lambda=0$, then

$$\mathbf{p}_{j+1}=(\mathbf{D}_\xi^\top\mathbf{B}^\top\mathbf{Q}^*\mathbf{B}\mathbf{D}_\xi)^{-1}\mathbf{D}_\xi^\top\mathbf{B}^\top\mathbf{Q}^*(\widetilde{\mathbf{x}}_1-\mathbf{u}-\mathbf{B}\mathbf{f}_\xi)\tag{183}$$

The difference is that $\mathbf{Q}^*$ appears to deal with the 0s on the diagonal of $\mathbf{Q}$ when taking the inverse.

6.2.11 B update equation for partially deterministic models when $\mathbf{B}^{(0)}$ is diagonal and not fixed

If $\mathbf{B}^{(0)}$ is diagonal and fixed, we can use the usual constrained update equation for $\mathbf{B}$. But if we wanted to estimate $\mathbf{B}^{(0)}$, the problem becomes difficult as outlined here. First we would write Ψ^+ in equation (172) as a function of $\boldsymbol{\beta}$ instead of $\mathbf{B}$. Note that $\mathbf{B}^{\diamondsuit}\boldsymbol{X}_0$ and $\mathbf{B}^{\sharp}\mathbf{u}$ are column vectors. We could use relation (74) to show that:

$$\begin{aligned}
&\mathbf{I}_q^{(0)}\mathbf{B}^{\diamondsuit}\mathbf{I}_q^{(0)}\boldsymbol{X}_0 = (\boldsymbol{X}_0^\top \otimes \mathbf{I})((\mathbf{f}_b^{(0)})^t + \mathbf{D}_b^{(0)}\boldsymbol{\beta}^t),\\
&\mathbf{I}_q^{(0)}\mathbf{B}^{\sharp}\mathbf{I}_q^{(0)}\mathbf{u} = (\mathbf{u}^\top \otimes \mathbf{I})((\mathbf{f}_b^{(0)})^\sharp + \mathbf{D}_b^{(0)}\boldsymbol{\beta}^\sharp),\\
&\quad \text{where } \mathbf{d}^t \equiv \begin{bmatrix} d_1^t\\ d_2^t\\ \dots\\ d_p^t \end{bmatrix}\\
&\quad \text{where } \mathbf{d}^\sharp \equiv \begin{bmatrix} d_1^t/(1-d_1)\\ d_2^t/(1-d_2)\\ \dots\\ d_p^t/(1-d_p) \end{bmatrix}
\end{aligned} \tag{184}$$

The terms $\mathbf{f}_b^{(0)}$ and $\mathbf{D}_b^{(0)}$ have the rows corresponding to $\mathrm{vec}(\mathbf{B}^+)$ zero'ed out.

The derivation I believe would proceed by taking the derivative of Ψ^+ with respect to $\boldsymbol{\beta}$. However we would end up with a polynomial in $\boldsymbol{\beta}$ because we will have the terms $\frac{\partial b^t}{\partial b}$ and $\frac{\partial b^t/(1-b)}{\partial b}$. where b denotes one of the diagonal elements in $\mathbf{B}^{(0)}$. That starts to look messy and there might be multiple solutions. Perhaps another day, I will solve that problem or come upon a more elegant solution. For now, I will side-step this problem and require that any $\mathbf{B}^{(0)}$ terms are fixed.

7 Implementation comments

The EM algorithm is a hill-climbing algorithm and like all hill-climbing algorithms it can get stuck on local maxima. There are a number approaches to doing a pre-search of the initial conditions space, but a brute force random Monte Carol search appears to work well (Biernacki et al., 2003). It is slow, but normally sufficient. In my experience, Monte Carlo initial conditions searches become important as the fraction of missing data in the data set increases. Certainly an initial conditions search should be done before reporting final estimates for an analysis. However in our[23] studies on the distributional properties of parameter estimates, we rarely found it necessary to do an initial conditions search.

The EM algorithm will quickly home in on parameter estimates that are close to the maximum, but once the values are close, the EM algorithm can slow to a crawl. Some researchers start with an EM algorithm to get close to the maximum-likelihood parameters and then switch to a quasi-Newton method for the final search. In many ecological applications, parameter estimates that differ by less than 3 decimal places are for all practical purposes the same. Thus we have not used the quasi-Newton final search.

Shumway and Stoffer (2006; chapter 6) imply in their discussion of the EM algorithm that both ξ and Λ can be estimated, though not simultaneously. Harvey (1989),

[23] "Our" and "we" in this section means work and papers by E. E. Holmes and E.J. Ward.

in contrast, discusses that there are only two allowable cases for the initial conditions: 1) fixed but unknown and 2) a initial condition set as a prior. In case 1, ξ is $\boldsymbol{x}_0$ (or $\boldsymbol{x}_1$) and is then estimated as a parameter; Λ is held fixed at 0. In case 2, ξ and Λ specify the mean and variance of $\boldsymbol{X}_0$ (or $\boldsymbol{X}_1$) respectively. Neither are estimated; instead, they are specified as part of the model.

As mentioned in the introduction, misspecification of the prior on $\boldsymbol{x}_0$ can have catastrophic and undetectable effects on your parameter estimates. For many MARSS models, you will never see this problem. However, if you are fitting models that imply a correlation structure between the hidden states (i.e. the variance-covariance matrix of the $\boldsymbol{X}$'s is not diagonal), then your prior can definitely create problems if it does not have the same correlation structure as that implied by your MLE model. A common default is to use a prior with a diagonal variance-covariance matrix. This can lead to serious problems if the implied variance-covariance of the $\boldsymbol{X}$'s is not diagonal. A diffuse prior does not get around this since it has a correlation structure also even if it has infinite variance.

One way you can detect that you have a problem is to start the EM algorithm at the outputs from a Newton-esque algorithm. If the EM estimates diverge and the likelihood drops, you have a problem. Here are a few suggestions for getting around the problem:

- Treat $\boldsymbol{x}_0$ as an estimated parameter and set $\mathbf{V}_0$=0. If the model is not stable going backwards in time, then treat $\boldsymbol{x}_1$ as the estimated parameter; this will allow the data to constrain the $\boldsymbol{x}_1$ estimate (since there is no data at $t = 0, \boldsymbol{x}_0$ has no data to constrain it).

- Try a diffuse prior, but first read the info in the KFAS R package about diffuse priors since MARSS uses the KFAS implementation. In particular, note that you will still be imposing an information on the correlation structure using a diffuse prior; whatever $\mathbf{V}_0$ you use is telling the algorithm what correlation structure to use. If there is a mismatch between the correlation structure in the prior and the correlation structure implied by the MLE model, you will not be escaping the prior problem. But sometimes you will know your implied correlation structure. For example, you may know that the $\boldsymbol{x}$'s are independent or you may be able to solve for the stationary distribution a priori if your stationary distribution is not a function of the parameters you are trying to estimate. Other times you are estimating a parameter that determines the correlation structure (like $\mathbf{B}$) and you will not know a priori what the correlation structure is.

In some cases, the update equation for one parameter needs other parameters. Technically, the Kalman filter/smoother should be run between each parameter update, however following Ghahramani and Hinton (1996) the default MARSS algorithm skips this step (unless the user sets `control$EMsafe=TRUE`) and each updated parameter is used for subsequent update equations.

8 MARSS R package

R code for the Kalman filter, Kalman smoother, and EM algorithm is provided as a separate R package, MARSS, available on CRAN (http://cran.r-project.org/web/packages/MARSS).

MARSS was developed by Elizabeth Holmes, Eric Ward and Kellie Wills and provides maximum-likelihood estimation and model-selection for both unconstrained and constrained MARSS models. The package contains a detailed user guide which shows various applications. In addition to model fitting via the EM algorithm, the package provides algorithms for bootstrapping, confidence intervals, auxiliary residuals, and model selection criteria.

References

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41(3-4):561–575.

Borman, S. (2009). *The Expectation Maximization Algorithm - A short tutorial*.

Ghahramani, Z. and Hinton, G. E. (1996). Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, Cambridge, UK.

Henderson, H. V. and Searle, S. R. (1979). Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics. *The Canadian Journal of Statistics*, 7(1):65–81.

Johnson, R. A. and Wichern, D. W. (2007). *Applied multivariate statistical analysis*. Prentice Hall, Upper Saddle River, NJ.

McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions*. John Wiley and Sons, Inc., Hoboken, NJ, 2nd edition.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear gaussian models. *Neural Computation*, 11:305–345.

Shumway, R. and Stoffer, D. (2006). *Time series analysis and its applications*. Springer-Science+Business Media, LLC, New York, New York, 2nd edition.

Shumway, R. H. and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4):253–264.

Wu, L. S.-Y., Pai, J. S., and Hosking, J. R. M. (1996). An algorithm for estimating parameters of state-space models. *Statistics and Probability Letters*, 28:99–106.

Zuur, A. F., Fryer, R. J., Jolliffe, I. T., Dekker, R., and Beukema, J. J. (2003). Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics*, 14(7):665–685.