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

July 31, 2011

Contents

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 Unpublished report.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 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 (a type of VAR state-space model), one should be congnizant that misspecification of the prior on the initial hidden states ($\boldsymbol{x}_0$) 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 is inconsistent with the distribution of $\boldsymbol{x}_0$ implied by the MLE model. This often happens when the model implies that $\boldsymbol{x}_0$ has covariance structure. But since you do not know the MLE parameters, you do not know the covariance structure of $\boldsymbol{x}_0$. Using a diffuse prior does not help since your diffuse prior still has some correlation

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

structure (often independence is being imposed). As mentioned above, often 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. 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 circumvent them.

1.1 The MARSS model

The linear MARSS model with a stochastic $\boldsymbol{x}_0$[2] is

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

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

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

Here $\boldsymbol{y}_t$ is a $n \times 1$ vector and $\boldsymbol{x}_t$ is a $m \times 1$ vector. The equation for $\boldsymbol{x}$ describes a multivariate Markov process or random walk. $\mathbf{B}$ is a $m \times m$ matrix, $\mathbf{u}$ is a $m \times 1$ matrix and $\mathbf{Q}$ is a $m \times m$ variance-covariance matrix. $\mathbf{Z}$ is a $n \times m$ matrix, $\mathbf{a}$ is a $n \times 1$ matrix and $\mathbf{R}$ is a $n \times n$ variance-covariance matrix. MARSS models are used in many different fields, and the meanings and names of these parameter matrices depend on the context and field in which the model is used.

Traditionally, $\boldsymbol{x}_t$ is treated as hidden, hence the name 'hidden Markov process' of which a MARSS model is a special type, and $\boldsymbol{y}_t$ is treated as (partially) observed. In this report, I follow this tradition, however it is certainly possible to treat both $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ as partially observed. If $\boldsymbol{x}_t$ is partially observed then the expectations used in the algorithm (section 5) would be computed conditioned on the partially observed $\boldsymbol{x}_t$.

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

[2] meaning $\boldsymbol{x}_0$ has a distribution rather than a fixed value

[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.

for $\boldsymbol{Y}_t$):[4]

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

where

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

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).

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

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 $\mathbf{V}_0$. The

[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 $f_{B,u,Q}(\boldsymbol{x}_t | \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1})$.

likelihood then is slightly different:

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

Note that in this case, $\boldsymbol{x}_0$ is no long a realization of a random variable $\boldsymbol{X}_0$; it is a fixed (but unknown) parameter. Note this is written as if all the $\mathbf{V}_0$ elements are 0 for the sake of clarity, however the MARSS package does require that all $\mathbf{V}_0$ 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 $\sum_1^T \frac{1}{2}\log|\mathbf{Q}|$.

Note that 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 $\widetilde{\mathbf{y}}_t = \mathrm{E}[\boldsymbol{Y}_t|\text{data}]$ and $\widetilde{\mathbf{y}\mathbf{x}}_t = \mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\text{data}]$ in the update equations (much like $\widetilde{\mathbf{x}}_t$ is kept in the equations) while Shumway and Stoffer replace these expectations involving $\boldsymbol{Y}_t$ by their values, which depend on whether there are missing values in the data. Section 5 shows how to compute the expectations involving $\boldsymbol{Y}_t$ when the data have or do not have missing values.

[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.

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):

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

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

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

Now, β_j appears in both the likelihood function and the $f()$ function, and the maximization is only over α.

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.

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 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 $\mathrm{E}_{\mathbf{XY}}[g(\boldsymbol{X}, \boldsymbol{Y})|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j]$. This is a multivariate conditional expectation because $\boldsymbol{X}, \boldsymbol{Y}$ is multivariate (a

$m \times n \times T$ vector). The function g that we are taking the expectation of is $\log \mathbf{L}(\boldsymbol{Y}, \boldsymbol{X}|\Theta)$. Note that $g()$ 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:

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

All the $\mathrm{E}[\ \ ]$ 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 $\boldsymbol{X}_0 = \boldsymbol{\xi}$ and the last two lines involving $\mathbf{V}_0$ are dropped.

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 long form. The goal is to find the Θ that maximize this expectation and those become 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).

$\mathrm{E}_X[g(\boldsymbol{X})] = \int\int\int g(\boldsymbol{x}) f(x_1, x_2, x_3) dx_1 dx_2 dx_3$
$\mathrm{E}_X[X_1] = \int\int\int x_1 f(x_1, x_2, x_3) dx_1 dx_2 dx_3 = \int x_1 f(x_1) dx_1 = \mathrm{E}[X_1]$
$\mathrm{E}_X[X_1 + X_2] = \mathrm{E}_X[X_1] + \mathrm{E}_X[X_2]$
$\mathrm{E}_X[X_1 + C] = \mathrm{E}_X[X_1] + C$
$\mathrm{E}_X[CX_1] = C\,\mathrm{E}_X[X_1]$
$\mathrm{E}_X[X_1 | X_1 = x_1] = x_1$
$\mathrm{E}_X[\boldsymbol{X} | \boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{x}$

2.2 The expectations used in the derivation

The following expectations will be needed in the derivation:

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

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:

$$\mathrm{var}[X] = \mathrm{E}[XX^\top] - \mathrm{E}[X]\,\mathrm{E}[X]^\top \tag{11}$$
$$\mathrm{cov}[X, Y] = \mathrm{E}[XY^\top] - \mathrm{E}[X]\,\mathrm{E}[Y]^\top. \tag{12}$$

Section 5 shows how to compute these expectations.

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.

To derive the update equations, one must find the Θ ($\Theta \neq \Theta_j$) that maximizes Ψ (equation 9) by partial differentiation of Ψ with 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 . . .) is

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

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

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

Note that the indexing is interchanged; $\partial \Psi / \partial b_{i,j} = \left[\partial \Psi / \partial \mathbf{B}\right]_{j,i}$. 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.

A number of derivatives of a scalar with respect to vectors and matrices will be needed in the derivation. 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, and $\mathbf{C}$ is a $n \times n$

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). Note, all the numerators in the differentials reduce to scalars.

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

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

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

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

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

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

matrix. Note, all the numerators in the differentials reduce to scalars. Both the vectorized and non-vectorized versions are shown, where

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

where $\mathbf{C}^{-1}$ 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 $\mathbf{C}$.

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 $\mathbf{u}$:

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

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

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

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[7] 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 symbols in 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 $\partial(\,\mathrm{E}[h(\boldsymbol{X}_t, \mathbf{B})])/\partial\mathbf{B} =$

[7] $\mathbf{Q}$ is a variance-covariance matrix and is invertable. $\mathbf{Q}^{-1}\mathbf{Q} = \mathbf{I}$, 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 $\mathbf{Q}^{-1} = (\mathbf{Q}^{-1})^\top$:

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

This can be reduced to

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

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 the expectations in equation (10). Solving for $\mathbf{B}$ and noting that $\widetilde{\mathbf{P}}_{t-1}$ is like a variance-covariance matrix and is invertable, 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 $\mathbf{B}$,

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

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 $\partial(\,\mathrm{E}[h(\boldsymbol{X}_t, \mathbf{Q})])/\partial\mathbf{Q} = \mathrm{E}[\partial(h(\boldsymbol{X}_t, \mathbf{Q}))/\partial\mathbf{Q}]$.

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

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 $-1/2$.

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

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

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

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

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

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.

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

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

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

This derivation immediately generalizes to the case where $\mathbf{Q}$ is a block di-

agonal 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,

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

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 $\mathbf{Q}$ 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.

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

The expectations around constants can be dropped[8]. Using relations (13) and (17) and using $\mathbf{R}^{-1} = (\mathbf{R}^{-1})^\top$, we have then

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

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

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

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 [9].

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

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

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

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

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

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

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

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

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

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

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.

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

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 $-1/2$.

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

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

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

We rewrite the partial derivative in terms of expectations:

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

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 $1/2$.

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

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

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

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,

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

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 $\boldsymbol{\xi}$ and $\mathbf{V}_0$ (unconstrained)

Shumway and Stoffer (2006) and Ghahramani and Hinton (1996) imply in their discussion of the EM algorithm that both $\boldsymbol{\xi}$ and $\mathbf{V}_0$ can be estimated (though not simultaneously). Harvey (1989), however, discusses that there are only two allowable cases: $\boldsymbol{x}_0$ is treated as fixed ($\mathbf{V}_0 = 0$) and equal to the unknown parameter $\boldsymbol{\xi}$ or $\boldsymbol{x}_0$ is treated as stochastic with a known mean $\boldsymbol{\xi}$ and variance $\mathbf{V}_0$. For completeness, we show here the update equation in the case of $\boldsymbol{x}_0$ stochastic with unknown mean $\boldsymbol{\xi}$ and variance $\mathbf{V}_0$ (a case that Harvey (1989) says is not consistent). If $\mathbf{V}_0$ is also treated as unknown, then the estimate of $\mathbf{V}_0$ at iteration j is used in place of $\mathbf{V}_0$. This latter case, where both $\boldsymbol{\xi}$ and $\mathbf{V}_0$ are estimated often gives researchers problems, but it might be feasible when your data and model are highly stationary so that $\boldsymbol{\xi}$ and $\mathbf{V}_0$ are well specified. Setting a fixed diffuse prior on $\boldsymbol{\xi}$ might also be reasonable, however, be aware that specifying a prior that is inconsistent with the data can cause problems. For example, if your multivariate data are correlated yet your $\mathbf{V}_0$ is diagonal and large, this can cause problems for the EM algorithm (likelihood does not increase and the algorithm wanders away from the maximum).

Let's start with the case where $\boldsymbol{\xi}$ is the mean of the distribution of $\boldsymbol{X}_0$. We can proceed as before and solve for the new $\boldsymbol{\xi}$ by minimizing Ψ. Take the derivative of Ψ with respect to $\boldsymbol{\xi}$. Terms not involving $\boldsymbol{\xi}$, equal 0 and drop out.

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

Using relations (13) and (17) and using $\mathbf{V}_1^{-1} = (\mathbf{V}_1^{-1})^\top$, we have

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

Pulling the parameters out of the expectations, we get

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

We then set the left side to zero, take the transpose, and cancel out $-1/2$ and $\mathbf{V}_1^{-1}$ (by noting that it is a variance-covariance matrix and is invertable).

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

Thus,

$$\boldsymbol{\xi}_{j+1} = \widetilde{\mathbf{x}}_0 \tag{58}$$

$\widetilde{\mathbf{x}}_0$ 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 $\mathrm{E}(\boldsymbol{X}_0|\boldsymbol{Y}_0 = \boldsymbol{y}_0) \equiv \boldsymbol{\xi}$ and $\mathrm{var}(\boldsymbol{X}_0\boldsymbol{X}_0^\top|\boldsymbol{Y}_0 = \boldsymbol{y}_0) \equiv \mathbf{V}_0$.

A similar set of steps gets us to the update equation for $\mathbf{V}_0$,

$$\mathbf{V}_{0,j+1} = \widetilde{\mathbf{V}}_0 \tag{59}$$

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

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

$$\begin{aligned}
\partial\Psi/\partial\boldsymbol{\xi} = -\frac{1}{2}\Big(&- \mathrm{E}[\partial(\boldsymbol{X}_1^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{\xi})/\partial\boldsymbol{\xi}] \\
&- \mathrm{E}[\partial((\mathbf{B}\boldsymbol{\xi})^\top\mathbf{Q}^{-1}\boldsymbol{X}_1)/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{\xi})^\top\mathbf{Q}^{-1}(\mathbf{B}\boldsymbol{\xi}))/\partial\boldsymbol{\xi}] \\
&+ \mathrm{E}[\partial((\mathbf{B}\boldsymbol{\xi})^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{\xi})/\partial\boldsymbol{\xi}]\Big) \\
= -\frac{1}{2}\Big(&- \mathrm{E}[\partial(\boldsymbol{X}_1^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{\xi})/\partial\boldsymbol{\xi}] \\
&- \mathrm{E}[\partial(\boldsymbol{\xi}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\boldsymbol{X}_1)/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial(\boldsymbol{\xi}^\top\mathbf{B}^\top\mathbf{Q}^{-1}(\mathbf{B}\boldsymbol{\xi}))/\partial\boldsymbol{\xi}] \\
&+ \mathrm{E}[\partial(\boldsymbol{\xi}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{\xi})/\partial\boldsymbol{\xi}]\Big)
\end{aligned} \tag{60}$$

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

$$\begin{aligned}
\partial\Psi/\partial\boldsymbol{\xi} = -\frac{1}{2}\Big(&- \mathrm{E}[\boldsymbol{X}_1]^\top\mathbf{Q}^{-1}\mathbf{B} - \mathrm{E}[\boldsymbol{X}_1]^\top\mathbf{Q}^{-1}\mathbf{B} \\
&+ 2\boldsymbol{\xi}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B} + \mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B} + \mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\Big)
\end{aligned} \tag{61}$$

This can be reduced to

$$\partial\Psi/\partial\boldsymbol{\xi} = \mathrm{E}[\boldsymbol{X}_1]^\top\mathbf{Q}^{-1}\mathbf{B} - \boldsymbol{\xi}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B} - \mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B} \tag{62}$$

To solve for $\boldsymbol{\xi}$, 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 $\boldsymbol{\xi}$. This step requires that this inverse exists.

$$\boldsymbol{\xi} = (\mathbf{B}^\top \mathbf{Q}^{-1} \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{Q}^{-1} (\,\mathrm{E}[\boldsymbol{X}_1] - \mathbf{u}) \tag{63}$$

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

$$\boldsymbol{\xi}_{j+1} = (\mathbf{B}^\top \mathbf{Q}^{-1} \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{Q}^{-1} (\widetilde{\mathbf{x}}_1 - \mathbf{u}) \tag{64}$$

Note that this is conceptually similar to using a generalized least squares estimate of $\boldsymbol{\xi}$ 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 (63) probably is not what you want to use. The estimate of $\boldsymbol{\xi}$ 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 $\boldsymbol{\xi}$ 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 $\boldsymbol{\xi}$ and $\mathbf{V}_0$, since you probably don't want to do that (if might start to get circular), you might try an iterative process to get decent $\boldsymbol{\xi}$ and $\mathbf{V}_0$ or try fixing $\boldsymbol{\xi}$ and estimating $\mathbf{V}_0$ (above). You can fix $\boldsymbol{\xi}$ 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 $\mathbf{a}$ term to account for non-zero mean data).

In some cases, the estimate of $\boldsymbol{x}_0$ from $\boldsymbol{x}_1$ will be highly sensitive to small changes in the parameters; this is particularly the case for certain $\mathbf{B}$ matrices, even if they are stationary. The result is that your $\boldsymbol{\xi}$ 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 $\boldsymbol{\xi}$ 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 $\boldsymbol{\xi}$. 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:

$$\begin{aligned}
\log \mathbf{L}(\boldsymbol{y}, \boldsymbol{x}|\Theta) = &- \sum_1^T \frac{1}{2} (\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a})^\top \mathbf{R}^{-1} (\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a}) - \sum_1^T \frac{1}{2} \log |\mathbf{R}| \\
&- \sum_2^T \frac{1}{2} (\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u})^\top \mathbf{Q}^{-1} (\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u}) - \sum_1^T \frac{1}{2} \log |\mathbf{Q}| \\
\boldsymbol{x}_1 &\equiv \boldsymbol{\xi}
\end{aligned} \tag{65}$$

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

$$
\begin{aligned}
\partial\Psi/\partial\boldsymbol{\xi} = -\frac{1}{2}\Big(&- \mathrm{E}[\partial(\boldsymbol{Y}_1^\top \mathbf{R}^{-1}\mathbf{Z}\boldsymbol{\xi})/\partial\boldsymbol{\xi}] \\
&- \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{\xi})^\top \mathbf{R}^{-1}\boldsymbol{Y}_1)/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{\xi})^\top \mathbf{R}^{-1}(\mathbf{Z}\boldsymbol{\xi}))/\partial\boldsymbol{\xi}] \\
&+ \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{\xi})^\top \mathbf{R}^{-1}\mathbf{a})/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial(\mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z}\boldsymbol{\xi})/\partial\boldsymbol{\xi}]\Big) \\
-\frac{1}{2}\Big(&- \mathrm{E}[\partial(\boldsymbol{X}_2^\top \mathbf{Q}^{-1}\mathbf{B}\boldsymbol{\xi})/\partial\boldsymbol{\xi}] \\
&- \mathrm{E}[\partial((\mathbf{B}\boldsymbol{\xi})^\top \mathbf{Q}^{-1}\boldsymbol{X}_2)/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{\xi})^\top \mathbf{Q}^{-1}(\mathbf{B}\boldsymbol{\xi}))/\partial\boldsymbol{\xi}] \\
&+ \mathrm{E}[\partial((\mathbf{B}\boldsymbol{\xi})^\top \mathbf{Q}^{-1}\mathbf{u})/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial(\mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\boldsymbol{\xi})/\partial\boldsymbol{\xi}]\Big)
\end{aligned}
\tag{66}
$$

Note that the second summation starts at $t = 2$ and $\boldsymbol{\xi}$ is $\boldsymbol{x}_1$ instead of $\boldsymbol{x}_0$. The update equation for $\boldsymbol{\xi}$ is then

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

$$
\begin{aligned}
\partial\Psi/\partial\boldsymbol{\xi} = -\frac{1}{2}\Big(&- \mathrm{E}[\boldsymbol{Y}_1]^\top \mathbf{R}^{-1}\mathbf{Z} - \mathrm{E}[\boldsymbol{Y}_1]^\top \mathbf{R}^{-1}\mathbf{Z} \\
&+ 2\boldsymbol{\xi}^\top \mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z}\Big) \\
-\frac{1}{2}\Big(&- \mathrm{E}[\boldsymbol{X}_2]^\top \mathbf{Q}^{-1}\mathbf{B} - \mathrm{E}[\boldsymbol{X}_2]^\top \mathbf{Q}^{-1}\mathbf{B} \\
&+ 2\boldsymbol{\xi}^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B} + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B} + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\Big)
\end{aligned}
\tag{67}
$$

This can be reduced to

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

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

$$
\boldsymbol{\xi} = (\mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B})^{-1}(\mathbf{Z}^\top \mathbf{R}^{-1}(\,\mathrm{E}[\boldsymbol{Y}_1] - \mathbf{a}) + \mathbf{B}^\top \mathbf{Q}^{-1}(\,\mathrm{E}[\boldsymbol{X}_2] - \mathbf{u}))
\tag{69}
$$

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

$$
\boldsymbol{\xi}_{j+1} = (\mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B})^{-1}(\mathbf{Z}^\top \mathbf{R}^{-1}(\widetilde{\mathbf{y}}_1 - \mathbf{a}) + \mathbf{B}^\top \mathbf{Q}^{-1}(\widetilde{\mathbf{x}}_2 - \mathbf{u}))
\tag{70}
$$

Note that using, say, $\widetilde{\mathbf{x}}_1$ output from the Kalman smoother would not work since $\mathbf{V}_1 = 0$ (in the case where $\boldsymbol{x}_1 \equiv \boldsymbol{\xi}$), and by definition $\widetilde{\mathbf{x}}_1 = \boldsymbol{\xi}$. Thus, you would never update your $\boldsymbol{\xi}$ estimate; it would simply stay fixed at whatever you started $\boldsymbol{\xi}$ in the EM algorithm.

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 elements are fixed values and some elements are shared (meaning they are forced to have the same value). One cannot simply use the elements from the unconstrained case for the free elements because the solution depends on the fixed values; those have to be included in the solution. One could always go through each matrix element one-by-one, but that would be very slow since the Kalman smoother would need to be run after updating each matrix element. Rather one would like to find a simultaneous solution for all the free elements in one's constrained parameter matrix.

Let's say we have some parameter matrix $\mathbf{M}$ (here $\mathbf{M}$ could be any of the parameters in the MARSS model) with fixed, shared and unshared elements:

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

The matrix $\mathbf{M}$ can be rewritten in terms of a fixed and free part, where in the fixed part all free elements are set to zero and in the free part all fixed elements are set to zero:

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

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

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

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 $\mathrm{vec}(\mathbf{M}_{\mathrm{free}})$. For example,

$$\mathrm{vec}(\mathbf{M}) = \begin{bmatrix} a \\ -1.2 \\ 0 \\ 0.9 \\ a \\ c \\ c \\ 0 \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ -1.2 \\ 0 \\ 0.9 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \mathbf{f} + \mathbf{D}\mathbf{m}$$

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 $\mathrm{vec}(\mathbf{M}) = \mathbf{f} + \mathbf{D}\mathbf{m}$. 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 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 $\mathbf{u} = \mathbf{f}_u + \mathbf{D}_u\boldsymbol{v}$, where $\boldsymbol{v}$ is the column vector of estimated parameters in $\mathbf{u}$. We then solve for $\partial\Psi/\partial\boldsymbol{v}$ by replacing $\mathbf{u}$ with $\mathbf{u} = \mathbf{f}_u + \mathbf{D}_u\boldsymbol{v}$ 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.

$$\begin{aligned} \partial\Psi/\partial\boldsymbol{v} = -\frac{1}{2}\sum_{t=1}^{T}\Big(& -\partial(\,\mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}(\mathbf{f}+\mathbf{D}\boldsymbol{v})])/\partial\boldsymbol{v} \\ & -\partial(\,\mathrm{E}[(\mathbf{f}+\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}\boldsymbol{X}_t])/\partial\boldsymbol{v} + \partial(\,\mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}(\mathbf{f}+\mathbf{D}\boldsymbol{v})])/\partial\boldsymbol{v} \\ & +\partial(\,\mathrm{E}[(\mathbf{f}+\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}])/\partial\boldsymbol{v} + \partial((\mathbf{f}+\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}(\mathbf{f}+\mathbf{D}\boldsymbol{v}))/\partial\boldsymbol{v}\Big) \end{aligned} \tag{79}$$

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

$$\begin{aligned} \partial\Psi/\partial\boldsymbol{v} = -\frac{1}{2}\sum_{t=1}^{T}\Big(& -\partial(\,\mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{v}])/\partial\boldsymbol{v} - \partial(\,\mathrm{E}[(\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}\boldsymbol{X}_t])/\partial\boldsymbol{v} \\ & +\partial(\,\mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{v}])/\partial\boldsymbol{v} + \partial(\,\mathrm{E}[(\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}])/\partial\boldsymbol{v} \\ & +\partial(\mathbf{f}^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{v})/\partial\boldsymbol{v} + \partial((\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}\mathbf{f})/\partial\boldsymbol{v} + \partial((\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{v})/\partial\boldsymbol{v}\Big) \end{aligned} \tag{80}$$

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.

$$\text{vec}(\mathbf{a}) = \text{vec}(\mathbf{a}^\top) = \mathbf{a}$$
The vec of a column vector (or its transpose) is itself. (71)

$$\text{vec}(\mathbf{Aa}) = (\mathbf{a}^\top \otimes \mathbf{I}_n)\,\text{vec}(\mathbf{A}) = \mathbf{Aa}$$
$\text{vec}(\mathbf{Aa}) = \mathbf{Aa}$ since $\mathbf{Aa}$ is itself an $m \times 1$ column vector. (72)

$$\text{vec}(\mathbf{AB}) = (\mathbf{I}_p \otimes \mathbf{A})\,\text{vec}(\mathbf{B}) = (\mathbf{B}^\top \otimes \mathbf{I}_n)\,\text{vec}(\mathbf{A}) \tag{73}$$

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\,\text{vec}(\mathbf{B}) \tag{74}$$

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC} \otimes \mathbf{BD}) \tag{75}$$

$$(\mathbf{a} \otimes \mathbf{I}_p)\mathbf{C} = (\mathbf{a} \otimes \mathbf{C})$$
$$\mathbf{C}(\mathbf{a}^\top \otimes \mathbf{I}_q) = (\mathbf{a}^\top \otimes \mathbf{C}) \tag{76}$$

$$(\mathbf{a} \otimes \mathbf{I}_p)\mathbf{C}(\mathbf{b}^\top \otimes \mathbf{I}_q) = (\mathbf{ab}^\top \otimes \mathbf{C}) \tag{77}$$

$$(\mathbf{a} \otimes \mathbf{a}) = \text{vec}(\mathbf{aa}^\top)$$
$$(\mathbf{a}^\top \otimes \mathbf{a}^\top) = (\mathbf{a} \otimes \mathbf{a})^\top = (\,\text{vec}(\mathbf{aa}^\top))^\top \tag{78}$$

Using the matrix differentiation relations in section 3.1, we get

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

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

$$\mathbf{0} = \sum_{t=1}^{T}\Big(\mathbf{D}^\top\mathbf{Q}^{-1}(\,\mathrm{E}[\boldsymbol{X}_t] - \mathbf{B}\,\mathrm{E}[\boldsymbol{X}_{t-1}] - \mathbf{f}) - \mathbf{D}^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{v}\Big) \tag{82}$$

Thus,

$$T\mathbf{D}^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{v} = \mathbf{D}^\top\mathbf{Q}^{-1}\sum_{t=1}^{T}\big(\,\mathrm{E}[\boldsymbol{X}_t] - \mathbf{B}\,\mathrm{E}[\boldsymbol{X}_{t-1}] - \mathbf{f}\big) \tag{83}$$

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

$$\boldsymbol{v}_{j+1} = \frac{1}{T}\big(\mathbf{D}_u^\top\mathbf{Q}^{-1}\mathbf{D}_u\big)^{-1}\mathbf{D}_u^\top\mathbf{Q}^{-1}\sum_{t=1}^{T}\big(\widetilde{\mathbf{x}}_t - \mathbf{B}\widetilde{\mathbf{x}}_{t-1} - \mathbf{f}_u\big) \tag{84}$$

and

$$\mathbf{u}_{j+1} = \mathbf{f}_u + \mathbf{D}_u\boldsymbol{v}_{j+1}, \tag{85}$$

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

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 $\mathbf{f}$)

$$\boldsymbol{\alpha}_{j+1} = \frac{1}{T}\big(\mathbf{D}_a^\top\mathbf{R}^{-1}\mathbf{D}_a\big)^{-1}\mathbf{D}_a^\top\mathbf{R}^{-1}\sum_{t=1}^{T}\big(\widetilde{\mathbf{y}}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{f}_a\big) \tag{86}$$

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

$$\mathbf{a}_{j+1} = \mathbf{f}_a + \mathbf{D}_a\boldsymbol{\alpha}_{j+1}, \tag{87}$$

If $\mathbf{R}$ is diagonal, this will reduce just updating the free elements in $\mathbf{a}$ using their values from the unconstrained update equation.

4.3 The general $\boldsymbol{\xi}$ update equation

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

$$\partial\Psi/\partial\mathbf{p} = \big(\widetilde{\mathbf{x}}_0^\top \mathbf{V}_0^{-1} - \boldsymbol{\xi}^\top \mathbf{V}_0^{-1}\big)\mathbf{D} \tag{88}$$

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

$$\mathbf{0} = \mathbf{D}^\top \big(\mathbf{V}_0^{-1}\widetilde{\mathbf{x}}_0 - \mathbf{V}_0^{-1}\mathbf{f} + \mathbf{V}_0^{-1}\mathbf{D}\mathbf{p}\big) \tag{89}$$

Thus,

$$\mathbf{p}_{j+1} = \big(\mathbf{D}_\xi^\top \mathbf{V}_0^{-1}\mathbf{D}_\xi\big)^{-1}\mathbf{D}_\xi^\top \mathbf{V}_0^{-1}(\widetilde{\mathbf{x}}_0 - \mathbf{f}_\xi) \tag{90}$$

and the new $\boldsymbol{\xi}$ is then,

$$\boldsymbol{\xi}_{j+1} = \mathbf{f}_\xi + \mathbf{D}_\xi \mathbf{p}_{j+1}, \tag{91}$$

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

$$\begin{aligned}
\partial\Psi/\partial\mathbf{p} = -\frac{1}{2}\Big(&- \mathrm{E}[\partial(\boldsymbol{X}_1^\top \mathbf{Q}^{-1}\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))/\partial\mathbf{p}] \\
&- \mathrm{E}[\partial((\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))^\top \mathbf{Q}^{-1}\boldsymbol{X}_1)/\partial\mathbf{p}] + \mathrm{E}[\partial((\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))^\top \mathbf{Q}^{-1}(\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p})))/\partial\mathbf{p}] \\
&+ \mathrm{E}[\partial((\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))^\top \mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{p}] + \mathrm{E}[\partial(\mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))/\partial\mathbf{p}]\Big) \\
= -\frac{1}{2}\Big(&- \mathrm{E}[\partial(\boldsymbol{X}_1^\top \mathbf{Q}^{-1}\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))/\partial\mathbf{p}] \\
&- \mathrm{E}[\partial((\mathbf{f}+\mathbf{D}\mathbf{p})^\top \mathbf{B}^\top \mathbf{Q}^{-1}\boldsymbol{X}_1)/\partial\mathbf{p}] + \mathrm{E}[\partial((\mathbf{f}+\mathbf{D}\mathbf{p})^\top \mathbf{B}^\top \mathbf{Q}^{-1}(\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p})))/\partial\mathbf{p}] \\
&+ \mathrm{E}[\partial((\mathbf{f}+\mathbf{D}\mathbf{p})^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{p}] + \mathrm{E}[\partial(\mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}(\mathbf{f}+\mathbf{D}\mathbf{p}))/\partial\mathbf{p}]\Big)
\end{aligned} \tag{92}$$

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

$$\begin{aligned}
\partial\Psi/\partial\mathbf{p} = -\frac{1}{2}\Big(&- \mathrm{E}[\boldsymbol{X}_1]^\top \mathbf{Q}^{-1}\mathbf{B}\mathbf{D} - \mathrm{E}[\boldsymbol{X}_1]^\top \mathbf{Q}^{-1}\mathbf{B}\mathbf{D} \\
&+ \mathbf{f}^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B}\mathbf{D} + \mathbf{f}^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B}\mathbf{D} \\
&+ 2\mathbf{p}^\top \mathbf{D}^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B}\mathbf{D} + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\mathbf{D} + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\mathbf{D}\Big)
\end{aligned} \tag{93}$$

This can be reduced to

$$\partial\Psi/\partial\mathbf{p} = \mathrm{E}[\boldsymbol{X}_1]^\top \mathbf{Q}^{-1}\mathbf{B}\mathbf{D} - \mathbf{f}^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B}\mathbf{D} - \mathbf{p}^\top \mathbf{D}^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B}\mathbf{D} - \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\mathbf{D} \tag{94}$$

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 $\mathbf{p}$. This step requires that this inverse exists.

$$\mathbf{p} = (\mathbf{D}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{B}^\top\mathbf{Q}^{-1}(\,\mathrm{E}[\boldsymbol{X}_1] - \mathbf{u} - \mathbf{B}\mathbf{f}) \tag{95}$$

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

$$\mathbf{p}_{j+1} = (\mathbf{D}_\xi^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B}\mathbf{D}_\xi)^{-1}\mathbf{D}_\xi^\top\mathbf{B}^\top\mathbf{Q}^{-1}(\widetilde{\mathbf{x}}_1 - \mathbf{u} - \mathbf{B}\mathbf{f}_\xi) \tag{96}$$

4.4 The general B update equation

The matrix $\mathbf{B}$ is rewritten as $\mathbf{B} = \mathbf{B}_{\text{fixed}} + \mathbf{B}_{\text{free}}$, thus $\mathrm{vec}(\mathbf{B}) = \mathbf{f}_b + \mathbf{D}_b\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the $p \times 1$ column vector of the p estimated values, $\mathbf{f}_b = \mathrm{vec}(\mathbf{B}_{\text{fixed}})$ and $\mathbf{D}_b\boldsymbol{\beta} = \mathrm{vec}(\mathbf{B}_{\text{free}})$. 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.

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\beta} = -\frac{1}{2}\sum_{t=1}^{T}\Big(&-\mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1})/\partial\boldsymbol{\beta}]\\ &-\mathrm{E}[\partial((\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\boldsymbol{X}_t)/\partial\boldsymbol{\beta}] + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}(\mathbf{B}\boldsymbol{X}_{t-1}))/\partial\boldsymbol{\beta}]\\ &+\mathrm{E}[\partial((\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\boldsymbol{\beta}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1})/\partial\boldsymbol{\beta}]\Big)\end{aligned} \tag{97}$$

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 (72) to show that:

$$\begin{aligned}\mathbf{B}\boldsymbol{X}_{t-1} &= \mathrm{vec}(\mathbf{B}\boldsymbol{X}_{t-1}) = \mathbf{K}_b\,\mathrm{vec}(\mathbf{B}) = \mathbf{K}_b(\mathbf{f}_b + \mathbf{D}_b\boldsymbol{\beta}),\\ &\text{where } \mathbf{K}_b = (\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I})\end{aligned} \tag{98}$$

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

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\beta} = -\frac{1}{2}\sum_{t=1}^{T}\Big(&-\mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))/\partial\boldsymbol{\beta}]\\ &-\mathrm{E}[\partial((\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))^\top\mathbf{Q}^{-1}\boldsymbol{X}_t)/\partial\boldsymbol{\beta}]\\ &+\mathrm{E}[\partial((\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))^\top\mathbf{Q}^{-1}\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))/\partial\boldsymbol{\beta}]\\ &+\mathrm{E}[\partial((\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\boldsymbol{\beta}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))/\partial\boldsymbol{\beta}]\Big)\end{aligned} \tag{99}$$

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

be simplified to

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\beta} = \\ -\frac{1}{2}\sum_{t=1}^{T}\Big(&-2\,\mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}\boldsymbol{\beta})/\partial\boldsymbol{\beta}] + 2\,\mathrm{E}[\partial((\mathbf{K}\mathbf{f})^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}\boldsymbol{\beta})/\partial\boldsymbol{\beta}] \\ &+ \mathrm{E}[\partial((\mathbf{K}\mathbf{D}\boldsymbol{\beta})^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}\boldsymbol{\beta})/\partial\boldsymbol{\beta}] + 2\,\mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}\boldsymbol{\beta})/\partial\boldsymbol{\beta}]\Big)\end{aligned} \tag{100}$$

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

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\beta} = \sum_{t=1}^{T}\Big(&\mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}] - \mathrm{E}[(\mathbf{K}\mathbf{f})^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}] \\ &- \mathrm{E}[\boldsymbol{\beta}^\top(\mathbf{K}\mathbf{D})^\top\mathbf{Q}^{-1}(\mathbf{K}\mathbf{D})] - \mathrm{E}[\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}]\Big)\end{aligned} \tag{101}$$

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

$$\begin{aligned}\mathbf{0} = \sum_{t=1}^{T}\Big(&\mathrm{E}[(\mathbf{K}\mathbf{D})^\top\mathbf{Q}^{-1}\boldsymbol{X}_t] - \mathrm{E}[(\mathbf{K}\mathbf{D})^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{f}] \\ &- \mathrm{E}[(\mathbf{K}\mathbf{D})^\top\mathbf{Q}^{-1}(\mathbf{K}\mathbf{D})]\boldsymbol{\beta} - \mathrm{E}[(\mathbf{K}\mathbf{D})^\top\mathbf{Q}^{-1}\mathbf{u}]\Big)\end{aligned} \tag{102}$$

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

$$\begin{aligned}\mathbf{0} = \\ \sum_{t=1}^{T}\Big(&\mathrm{E}[((\boldsymbol{X}_{t-1}^\top\otimes\mathbf{I})\mathbf{D})^\top\mathbf{Q}^{-1}\boldsymbol{X}_t] - \mathrm{E}[((\boldsymbol{X}_{t-1}^\top\otimes\mathbf{I})\mathbf{D})^\top\mathbf{Q}^{-1}(\boldsymbol{X}_{t-1}^\top\otimes\mathbf{I})\mathbf{f}] \\ &- \mathrm{E}[((\boldsymbol{X}_{t-1}^\top\otimes\mathbf{I})\mathbf{D})^\top\mathbf{Q}^{-1}(\boldsymbol{X}_{t-1}^\top\otimes\mathbf{I})\mathbf{D}]\boldsymbol{\beta} - \mathrm{E}[((\boldsymbol{X}_{t-1}^\top\otimes\mathbf{I})\mathbf{D})^\top\mathbf{Q}^{-1}\mathbf{u}]\Big)\end{aligned} \tag{103}$$

This looks daunting, but using relation (72) and noting that $(\mathbf{A}\otimes\mathbf{B})^\top = (\mathbf{A}^\top\otimes\mathbf{B}^\top)$, we can simplify equation (103) using the following:

$$\begin{aligned}(\boldsymbol{X}_{t-1}^\top\otimes\mathbf{I})\mathbf{Q}^{-1}\mathbf{u} &= (\boldsymbol{X}_{t-1}\otimes\mathbf{I})\mathbf{Q}^{-1}\mathbf{u} \\ &= (\boldsymbol{X}_{t-1}\otimes\mathbf{I})\,\mathrm{vec}(\mathbf{Q}^{-1}\mathbf{u}),\ \text{because } \mathbf{Q}^{-1}\mathbf{u} \text{ is a column vector} \\ &= \mathrm{vec}(\mathbf{Q}^{-1}\mathbf{u}(\boldsymbol{X}_{t-1})^\top),\ \text{using relation (72)}\end{aligned}$$

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 (77):

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

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 (103), we get

$$\begin{aligned} \mathbf{0} = \sum_{t=1}^{T} \Big(& \mathrm{E}[\mathbf{D}^\top \operatorname{vec}(\mathbf{Q}^{-1}\boldsymbol{X}_t\boldsymbol{X}_{t-1}^\top)] - \mathrm{E}[\mathbf{D}^\top (\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top \otimes \mathbf{Q}^{-1})\mathbf{f}] \\ & - \mathrm{E}[\mathbf{D}^\top (\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top \otimes \mathbf{Q}^{-1})\mathbf{D}]\boldsymbol{\beta} - \mathrm{E}[\mathbf{D}^\top \operatorname{vec}(\mathbf{Q}^{-1}\mathbf{u}\boldsymbol{X}_{t-1}^\top)] \Big) \end{aligned} \tag{104}$$

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

$$\begin{aligned} \mathbf{0} = \sum_{t=1}^{T} \Big(& \mathbf{D}^\top \operatorname{vec}(\mathbf{Q}^{-1}\widetilde{\mathbf{P}}_{t,t-1}) - \mathbf{D}^\top (\widetilde{\mathbf{P}}_{t-1} \otimes \mathbf{Q}^{-1})\mathbf{f} \\ & - \mathbf{D}^\top (\widetilde{\mathbf{P}}_{t-1} \otimes \mathbf{Q}^{-1})\mathbf{D}\boldsymbol{\beta} - \mathbf{D}^\top \operatorname{vec}(\mathbf{Q}^{-1}\mathbf{u}(\widetilde{\mathbf{x}}_{t-1})^\top) \Big) \end{aligned} \tag{105}$$

Solving for $\boldsymbol{\beta}$,

$$\begin{aligned} \boldsymbol{\beta}_{j+1} = & \Big(\sum_{t=1}^{T} \mathbf{D}_b^\top (\widetilde{\mathbf{P}}_{t-1} \otimes \mathbf{Q}^{-1})\mathbf{D}_b \Big)^{-1} \mathbf{D}_b^\top \Big(\sum_{t=1}^{T} \big(\operatorname{vec}(\mathbf{Q}^{-1}\widetilde{\mathbf{P}}_{t,t-1}) \\ & - (\widetilde{\mathbf{P}}_{t-1} \otimes \mathbf{Q}^{-1})\mathbf{f}_b - \operatorname{vec}(\mathbf{Q}^{-1}\mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top) \big) \Big) \end{aligned} \tag{106}$$

This requires that $(\mathbf{D}_b^\top (\widetilde{\mathbf{P}}_{t-1} \otimes \mathbf{Q}^{-1})\mathbf{D}_b)$ is invertable, which it is because it is a $p \times p$ diagonal matrix with only non-zero values on the diagonal.

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

$$\operatorname{vec}(\mathbf{B}_{j+1}) = \mathbf{f}_b + \mathbf{D}_b\boldsymbol{\beta}_{j+1}, \tag{107}$$

When there are no fixed or shared values, $\mathbf{B}_{\text{fixed}}$ equals zero and $\mathbf{D}_b$ equals an identity matrix. Equation (106) 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 then factored out.

4.5 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 $\operatorname{vec}(\mathbf{Z}) = \mathbf{f}_z + \mathbf{D}_z\boldsymbol{\zeta}$, where $\boldsymbol{\zeta}$ is the column vector of the p estimated

values, $\mathbf{f}_z = \text{vec}(\mathbf{Z}_{\text{fixed}})$ and $\mathbf{D}_z\boldsymbol{\zeta} = \text{vec}(\mathbf{Z}_{\text{free}})$. With the z subscript dropped off $\mathbf{D}$ and $\mathbf{f}$, the update equation for $\mathbf{Z}$ is

$$\begin{aligned}\boldsymbol{\zeta}_{j+1} = \Bigg(\sum_{t=1}^{T}(\mathbf{D}_z^\top(\widetilde{\mathbf{P}}_t \otimes \mathbf{R}^{-1})\mathbf{D}_z)\Bigg)^{-1}\mathbf{D}_z^\top\Bigg(\sum_{t=1}^{T}\big(\text{vec}(\mathbf{R}^{-1}\widetilde{\mathbf{y}\mathbf{x}}_t) \\ - (\widetilde{\mathbf{P}}_t \otimes \mathbf{R}^{-1})\mathbf{f}_z - \text{vec}(\mathbf{R}^{-1}\mathbf{a}\widetilde{\mathbf{x}}_t^\top)\big)\Bigg)\end{aligned} \tag{108}$$

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

$$\text{vec}(\mathbf{Z}_{j+1}) = \mathbf{f}_z + \mathbf{D}_z\boldsymbol{\zeta}_{j+1} \tag{109}$$

4.6 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 $\text{vec}(\mathbf{Q})$. 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 4.6.5). 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 is

$$\begin{aligned}\boldsymbol{q}_{j+1} &= \frac{1}{T}(\mathbf{D}_q^\top\mathbf{D}_q)^{-1}\mathbf{D}_q^\top\,\text{vec}(\mathbf{S}) \\ \text{vec}(\mathbf{Q})_{j+1} &= \mathbf{f}_q + \mathbf{D}_q\boldsymbol{q}_{j+1} \\ \text{where } \mathbf{S} &= \sum_{t=1}^{T}\big(\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}\widetilde{\mathbf{x}}_t^\top + \\ &\qquad \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\big)\end{aligned} \tag{110}$$

The matrices $\mathbf{f}_q$, $\mathbf{D}_q$, and $\boldsymbol{q}$ have their standard definitions (section 4). The vec of $\mathbf{Q}$ is written in the form of $\text{vec}(\mathbf{Q}) = \mathbf{f}_q + \mathbf{D}_q\boldsymbol{q}$, 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.

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.

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

Let $\mathbf{Q}$ be some 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 $\text{vec}(\mathbf{Q}^{-1}) = \mathbf{f}_q^* + \mathbf{D}_q \boldsymbol{q}^*$, 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.

$$\begin{aligned} \partial\Psi/\partial\boldsymbol{q}^* = -\frac{1}{2}\sum_{t=1}^{T} \partial\Big(& \text{E}[\boldsymbol{X}_t^\top \mathbf{Q}^{-1} \boldsymbol{X}_t] - \text{E}[\boldsymbol{X}_t^\top \mathbf{Q}^{-1} \mathbf{B} \boldsymbol{X}_{t-1}] \\ & - \text{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top \mathbf{Q}^{-1} \boldsymbol{X}_t] - \text{E}[\boldsymbol{X}_t^\top \mathbf{Q}^{-1} \mathbf{u}] \\ & - \text{E}[\mathbf{u}^\top \mathbf{Q}^{-1} \boldsymbol{X}_t] + \text{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top \mathbf{Q}^{-1} \mathbf{B} \boldsymbol{X}_{t-1}] \\ & + \text{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top \mathbf{Q}^{-1} \mathbf{u}] + \text{E}[\mathbf{u}^\top \mathbf{Q}^{-1} \mathbf{B} \boldsymbol{X}_{t-1}] + \mathbf{u}^\top \mathbf{Q}^{-1} \mathbf{u} \Big) / \partial \boldsymbol{q}^* \\ & - \partial\big(\frac{T}{2} \log |\mathbf{Q}|\big) / \partial \boldsymbol{q}^* \end{aligned} \tag{111}$$

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.[11]

$$
\begin{aligned}
\partial\Psi/\partial\boldsymbol{q}^{*} &= -\frac{1}{2}\sum_{t=1}^{T}\partial\Big(\mathrm{E}[\boldsymbol{X}_t^{\top}\otimes\boldsymbol{X}_t^{\top}] - \mathrm{E}[\boldsymbol{X}_t^{\top}\otimes(\mathbf{B}\boldsymbol{X}_{t-1})^{\top}] - \mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^{\top}\otimes\boldsymbol{X}_t^{\top}] \\
&\quad - \mathrm{E}[\boldsymbol{X}_t^{\top}\otimes\mathbf{u}^{\top}] - \mathrm{E}[\mathbf{u}^{\top}\otimes\boldsymbol{X}_t^{\top}] + \mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^{\top}\otimes(\mathbf{B}\boldsymbol{X}_{t-1})^{\top}] \\
&\quad + \mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^{\top}\otimes\mathbf{u}^{\top}] + \mathrm{E}[\mathbf{u}^{\top}\otimes(\mathbf{B}\boldsymbol{X}_{t-1})^{\top}] + (\mathbf{u}^{\top}\otimes\mathbf{u}^{\top})\Big)\operatorname{vec}(\mathbf{Q}^{-1})/\partial\boldsymbol{q}^{*} \\
&- \partial\Big(\frac{T}{2}\log|\mathbf{Q}|\Big)/\partial\boldsymbol{q}^{*} \\
&= -\frac{1}{2}\sum_{t=1}^{T}\partial\big(\operatorname{vec}(\mathbf{S})^{\top}\big)\operatorname{vec}(\mathbf{Q}^{-1})/\partial\boldsymbol{q}^{*} + \partial\big(\frac{T}{2}\log|\mathbf{Q}^{-1}|\big)/\partial\boldsymbol{q}^{*} \\
\text{where } \mathbf{S} &= \sum_{t=1}^{T}\big(\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^{\top} - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^{\top} - \mathbf{u}\widetilde{\mathbf{x}}_t^{\top} + \\
&\quad \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^{\top} + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^{\top} + \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^{\top}\mathbf{B}^{\top} + \mathbf{u}\mathbf{u}^{\top}\big)
\end{aligned}
\tag{112}
$$

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,

$$
\begin{aligned}
\partial\Psi/\partial\boldsymbol{q}^{*} = -\Big(&\frac{1}{2}\operatorname{vec}(\mathbf{S})^{\top}(\mathbf{f}^{*} + \mathbf{D}\boldsymbol{q}^{*}) \\
&- \frac{T}{2}(\log(f_1^{*}) + \log(f_2^{*})...k\log(q_1^{*}) + l\log(q_2^{*})...)\Big)/\partial\boldsymbol{q}^{*}
\end{aligned}
\tag{113}
$$

where k is the number of times q_1 appears on the diagonal of $\mathbf{Q}$ and l is the number of times q_2 appears, etc. Taking the derivatives,

$$
\begin{aligned}
\partial\Psi/\partial\boldsymbol{q}^{*} &== \frac{1}{2}\mathbf{D}^{\top}\operatorname{vec}(\mathbf{S}) - \frac{T}{2}(\log(f_1^{*}) + ...k\log(q_1^{*}) + l\log(q_2^{*})...)/\partial\boldsymbol{q}^{*} \\
&= \frac{1}{2}\mathbf{D}^{\top}\operatorname{vec}(\mathbf{S}) - \frac{T}{2}\mathbf{D}^{\top}\mathbf{D}\boldsymbol{q}
\end{aligned}
\tag{114}
$$

$\mathbf{D}^{\top}\mathbf{D}$ is a $p\times p$ matrix with k, l, etc. along the diagonal and thus is invertable; 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 $\mathbf{Q}$:

$$
\begin{aligned}
\boldsymbol{q}_{j+1} &= \frac{1}{T}(\mathbf{D}^{\top}\mathbf{D})^{-1}\mathbf{D}^{\top}\operatorname{vec}(\mathbf{S}) \\
\operatorname{vec}(\mathbf{Q})_{j+1} &= \mathbf{f} + \mathbf{D}\boldsymbol{q}_{j+1}
\end{aligned}
\tag{115}
$$

where $\mathbf{S}$ is defined in equation (112) and, as usual, $\mathbf{D}$ and $\mathbf{f}$ are the parameter specific matrices. In this case, $\mathbf{D} = \mathbf{D}_q$ and $\mathbf{f} = \mathbf{f}_q$.

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

4.6.2 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 $\log|\mathbf{Q}^{-1}|$:

$$\partial\Psi/\partial\boldsymbol{q}^{*} = \partial\Big(-\frac{1}{2}\sum_{t=1}^{T}\big(\operatorname{vec}(\mathbf{S})^{\top}\big)\operatorname{vec}(\mathbf{Q}^{-1}) + \frac{T}{2}\log|\mathbf{Q}^{-1}|\Big)/\partial\boldsymbol{q}^{*} \tag{116}$$

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:

$$\partial\Psi/\partial\boldsymbol{q}^{*} = \frac{1}{2}\mathbf{D}^{\top}\operatorname{vec}(\mathbf{S}) - \frac{T}{2}\mathbf{D}^{\top}\mathbf{D}\boldsymbol{q} \tag{117}$$

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 (115) in which $\mathbf{f}_q = 0$ since there are no fixed values.

4.6.3 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.

4.6.4 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 ($\mathbb{E}$).

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

In it's most general form, $\mathbf{Q}$ 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).

4.7 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

$$\begin{aligned} \boldsymbol{\rho}_{j+1} &= \frac{1}{T}(\mathbf{D}_r^\top \mathbf{D}_r)^{-1}\mathbf{D}_r^\top \operatorname{vec}\Big(\sum_{t=1}^{T} \mathbf{R}_{t,j+1}\Big) \\ \operatorname{vec}(\mathbf{R})_{j+1} &= \mathbf{f}_r + \mathbf{D}_r\boldsymbol{\rho}_{j+1} \end{aligned} \tag{118}$$

The $\mathbf{R}_{t,j+1}$ used at time step t in equation (118) is the term that appears in the summation in the unconstrained update equation with no missing values (equation 52):

$$\begin{aligned} \mathbf{R}_{t,j+1} = \Big(&\widetilde{\mathbf{O}}_t - \widetilde{\mathbf{y}\mathbf{x}}_t\mathbf{Z}^\top - \mathbf{Z}\widetilde{\mathbf{y}\mathbf{x}}_t^\top - \widetilde{\mathbf{y}}_t\mathbf{a}^\top - \mathbf{a}\widetilde{\mathbf{y}}_t^\top \\ &+ \mathbf{Z}\widetilde{\mathbf{P}}_t\mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{x}}_t\mathbf{a}^\top + \mathbf{a}\widetilde{\mathbf{x}}_t^\top\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\Big) \end{aligned} \tag{119}$$

5 Computing the expectations

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 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 observations at time t. If some elements are missing, that means some elements are equal to NA (or some other missing values marker):

$$\boldsymbol{y}_t = \begin{bmatrix} y_1 \\ NA \\ y_3 \\ y_4 \\ NA \\ y_6 \end{bmatrix} \tag{120}$$

We denote the observations as $\boldsymbol{y}_t(1)$ and the NAs 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 $\boldsymbol{\Omega}_t^{(1)}$ be the matrix that extracts only $\boldsymbol{Y}_t(1)$ from $\boldsymbol{Y}_t$ and $\boldsymbol{\Omega}_t(2)$ be the matrix that extracts only $\boldsymbol{Y}_t(2)$. For the example above,

$$\begin{aligned} \boldsymbol{Y}_t(1) &= \boldsymbol{\Omega}_t^{(1)} \boldsymbol{Y}_t, \quad \boldsymbol{\Omega}_t^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \\ \boldsymbol{Y}_t(2) &= \boldsymbol{\Omega}_t^{(2)} \boldsymbol{Y}_t, \quad \boldsymbol{\Omega}_t^{(2)} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{aligned} \tag{121}$$

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,

$$\begin{aligned} \mathbf{I}_t^{(1)} &= (\boldsymbol{\Omega}_t^{(1)})^\top \boldsymbol{\Omega}_t^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \\ \mathbf{I}_t^{(2)} &= (\boldsymbol{\Omega}_t^{(2)})^\top \boldsymbol{\Omega}_t^{(2)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \end{aligned} \tag{122}$$

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.

$$\widetilde{\mathbf{x}}_t = \mathrm{E}[\boldsymbol{X}_t] \tag{123a}$$

$$\widetilde{\mathbf{V}}_t = \mathrm{var}[\boldsymbol{X}_t] \tag{123b}$$

$$\widetilde{\mathbf{V}}_{t,t-1} = \mathrm{cov}[\boldsymbol{X}_t, \boldsymbol{X}_{t-1}] \tag{123c}$$

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

$$\widetilde{\mathbf{P}}_t = \mathrm{E}[\boldsymbol{X}_t \boldsymbol{X}_t^\top] = \widetilde{\mathbf{V}}_t + \widetilde{\mathbf{x}}_t \widetilde{\mathbf{x}}_t^\top \tag{123d}$$

$$\widetilde{\mathbf{P}}_{t,t-1} = \mathrm{E}[\boldsymbol{X}_t \boldsymbol{X}_{t-1}^\top] = \widetilde{\mathbf{V}}_{t,t-1} + \widetilde{\mathbf{x}}_t \widetilde{\mathbf{x}}_{t-1}^\top \tag{123e}$$

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,

$$\boldsymbol{y}_t = \begin{bmatrix} y_1 \\ 0 \\ y_3 \\ y_4 \\ 0 \\ y_6 \end{bmatrix}, \quad \mathbf{a}_t = \begin{bmatrix} a_1 \\ 0 \\ a_3 \\ a_4 \\ 0 \\ a_6 \end{bmatrix}, \quad \mathbf{Z}_t = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots \\ 0 & 0 & \cdots \\ z_{3,1} & z_{3,2} & \cdots \\ z_{4,1} & z_{4,2} & \cdots \\ 0 & 0 & \cdots \\ z_{6,1} & z_{6,2} & \cdots \end{bmatrix} \tag{124}$$

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,

$$\mathbf{R} = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} & r_{1,5} & r_{1,6} \\ r_{2,1} & r_{2,2} & r_{2,3} & r_{2,4} & r_{2,5} & r_{2,6} \\ r_{3,1} & r_{3,2} & r_{3,3} & r_{3,4} & r_{3,5} & r_{3,6} \\ r_{4,1} & r_{4,2} & r_{4,3} & r_{4,4} & r_{4,5} & r_{4,6} \\ r_{5,1} & r_{5,2} & r_{5,3} & r_{5,4} & r_{5,5} & r_{5,6} \\ r_{6,1} & r_{6,2} & r_{6,3} & r_{6,4} & r_{6,5} & r_{6,6} \end{bmatrix} \tag{125}$$

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:

$$\mathbf{R}_t = \begin{bmatrix} r_{1,1} & 0 & r_{1,3} & r_{1,4} & 0 & r_{1,6} \\ 0 & r_{2,2} & 0 & 0 & r_{2,5} & 0 \\ r_{3,1} & 0 & r_{3,3} & r_{3,4} & 0 & r_{3,6} \\ r_{4,1} & 0 & r_{4,3} & r_{4,4} & 0 & r_{4,6} \\ 0 & r_{5,2} & 0 & 0 & r_{5,5} & 0 \\ r_{6,1} & 0 & r_{6,3} & r_{6,4} & 0 & r_{6,6} \end{bmatrix} \tag{126}$$

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

$$\begin{aligned} \boldsymbol{y}_t &= \mathbf{I}_t^{(1)} \boldsymbol{y}_t \\ \mathbf{a}_t &= \mathbf{I}_t^{(1)} \mathbf{a} \\ \mathbf{Z}_t &= \mathbf{I}_t^{(1)} \mathbf{Z} \\ \mathbf{R}_t &= \mathbf{I}_t^{(1)} \mathbf{R} \mathbf{I}_t^{(1)} + \mathbf{I}_t^{(2)} \mathbf{R} \mathbf{I}_t^{(2)} \end{aligned} \tag{127}$$

$\mathbf{a}_t$, $\mathbf{Z}_t$ 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[12] and then the expectations are given by the following equations. The derivations for these equations are given in the subsections to follow.

$$\widetilde{\mathbf{y}}_t = \mathrm{E}[\boldsymbol{Y}_t] = \boldsymbol{y}_t - \nabla_t(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a}) \tag{128a}$$

$$\widetilde{\mathbf{O}}_t = \mathrm{E}[\boldsymbol{Y}_t \boldsymbol{Y}_t^\top] = \mathbf{I}_t^{(2)}(\nabla_t \mathbf{R} + \nabla_t \mathbf{Z} \widetilde{\mathbf{V}}_t \mathbf{Z}^\top \nabla_t^\top)\mathbf{I}_t^{(2)} + \widetilde{\mathbf{y}}_t \widetilde{\mathbf{y}}_t^\top \tag{128b}$$

$$\widetilde{\mathbf{y}\mathbf{x}}_t = \mathrm{E}[\boldsymbol{Y}_t \boldsymbol{X}_t^\top] = \nabla_t \mathbf{Z} \widetilde{\mathbf{V}}_t + \widetilde{\mathbf{y}}_t \widetilde{\mathbf{x}}_t^\top \tag{128c}$$

$$\widetilde{\mathbf{y}\mathbf{x}}_{t,t-1} = \mathrm{E}[\boldsymbol{Y}_t \boldsymbol{X}_{t-1}^\top] = \nabla_t \mathbf{Z} \widetilde{\mathbf{V}}_{t,t-1} + \widetilde{\mathbf{y}}_t \widetilde{\mathbf{x}}_{t-1}^\top \tag{128d}$$

$$\text{where } \nabla_t = \mathbf{I} - \mathbf{R}(\boldsymbol{\Omega}_t^{(1)})^\top (\boldsymbol{\Omega}_t^{(1)} \mathbf{R} (\boldsymbol{\Omega}_t^{(1)})^\top)^{-1} \boldsymbol{\Omega}_t^{(1)} \tag{128e}$$

$$\text{and } \mathbf{I}_t^{(2)} = (\boldsymbol{\Omega}_t^{(2)})^\top \boldsymbol{\Omega}_t^{(2)} \tag{128f}$$

If $\boldsymbol{y}_t$ is all missing, $\boldsymbol{\Omega}_t^{(1)}$ is a $0 \times n$ matrix, and we define $(\boldsymbol{\Omega}_t^{(1)})^\top (\boldsymbol{\Omega}_t^{(1)} \mathbf{R} (\boldsymbol{\Omega}_t^{(1)})^\top)^{-1} \boldsymbol{\Omega}_t^{(1)}$ to be a $n \times n$ matrix of zeros. If $\mathbf{R}$ is diagonal, then $\mathbf{R}(\boldsymbol{\Omega}_t^{(1)})^\top (\boldsymbol{\Omega}_t^{(1)} \mathbf{R} (\boldsymbol{\Omega}_t^{(1)})^\top)^{-1} \boldsymbol{\Omega}_t^{(1)} = \mathbf{I}_t^{(1)}$ 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.

[12] 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.

If there are zeros on the diagonal of $\mathbf{R}$ (section 6), the definition of Δ_t is changed slightly. 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

$$\nabla_t = \mathbf{I} - \mathbf{R}(\mho_t^{(r)})^\top (\mho_t^{(r)} \mathbf{R} (\mho_t^{(r)})^\top)^{-1} \mho_t^{(r)} \tag{129}$$

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

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

$$\mathrm{E}[\boldsymbol{Y}_t | \boldsymbol{Y}(1) = \boldsymbol{y}(1)] = \mathrm{E}[\boldsymbol{Y}_t | \boldsymbol{Y}_t = \boldsymbol{y}_t] = \boldsymbol{y}_t \tag{130}$$

If there are no observed values, then

$$\mathrm{E}[\boldsymbol{Y}_t | \boldsymbol{Y}(1) = \boldsymbol{y}(1)] = \mathrm{E}[\boldsymbol{Y}_t] = \mathrm{E}[\mathbf{Z}\boldsymbol{X}_t + \mathbf{a} + \mathbf{V}_t] = \mathbf{Z}\widetilde{\mathbf{x}}_t + \mathbf{a} \tag{131}$$

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):

$$\text{If, } \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \tag{132}$$

Then,

$$\begin{aligned} &(Y_1 | Y_1 = y_1) = y_1, \quad \text{and} \\ &(Y_2 | Y_1 = y_1) \sim \mathrm{MVN}(\bar{\mu}, \bar{\Sigma}), \quad \text{where} \\ &\qquad \bar{\mu} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_1) \\ &\qquad \bar{\Sigma} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned} \tag{133}$$

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

$$\begin{bmatrix} \boldsymbol{Y}_t(1) | \boldsymbol{X}_t = \boldsymbol{x}_t \\ \boldsymbol{Y}_t(2) | \boldsymbol{X}_t = \boldsymbol{x}_t \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} \boldsymbol{\Omega}_t^{(1)}(\mathbf{Z}\boldsymbol{x}_t + \mathbf{a}) \\ \boldsymbol{\Omega}_t^{(2)}(\mathbf{Z}\boldsymbol{x}_t + \mathbf{a}) \end{bmatrix}, \begin{bmatrix} \mathbf{R}_{t,11} & \mathbf{R}_{t,12} \\ \mathbf{R}_{t,21} & \mathbf{R}_{t,22} \end{bmatrix} \right) \tag{134}$$

Thus,

$$\begin{aligned} &(\boldsymbol{Y}_t(1) | \boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1), \boldsymbol{X}_t = \boldsymbol{x}_t) = \boldsymbol{\Omega}_t^{(1)} \boldsymbol{y}_t \quad \text{and} \\ &(\boldsymbol{Y}_t(2) | \boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1), \boldsymbol{X}_t = \boldsymbol{x}_t) \sim \mathrm{MVN}(\ddot{\mu}, \ddot{\Sigma}) \quad \text{where} \\ &\qquad \ddot{\mu} = \boldsymbol{\Omega}_t^{(2)}(\mathbf{Z}\boldsymbol{x}_t + \mathbf{a}) + \mathbf{R}_{t,21}(\mathbf{R}_{t,11})^{-1}\boldsymbol{\Omega}_t^{(1)}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a}) \\ &\qquad \ddot{\Sigma} = \mathbf{R}_{t,22} - \mathbf{R}_{t,21}(\mathbf{R}_{t,11})^{-1}\mathbf{R}_{t,12} \end{aligned} \tag{135}$$

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:

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

From this and the distributions in equation (135), we can write down $\widetilde{\mathbf{y}}_t = \mathrm{E}[\boldsymbol{Y}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j]$:

$$\begin{aligned}
\widetilde{\mathbf{y}}_t &= \mathrm{E}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1)] \\
&= \int_{\boldsymbol{x}_t}\int_{\boldsymbol{y}_t} \boldsymbol{y}_t f(\boldsymbol{y}_t|\boldsymbol{y}_t(1), \boldsymbol{x}_t) d\boldsymbol{y}_t f(\boldsymbol{x}_t) d\boldsymbol{x}_t \\
&= \mathrm{E}_X[\,\mathrm{E}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1), \boldsymbol{X}_t = \boldsymbol{x}_t]] \\
&= \mathrm{E}_X[\boldsymbol{y}_t - \nabla_t(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{X}_t - \mathbf{a})] \\
&= \boldsymbol{y}_t - \nabla_t(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a}) \\
\text{where } \nabla_t &= \mathbf{I} - \mathbf{R}(\boldsymbol{\Omega}_t^{(1)})^\top(\mathbf{R}_{t,11})^{-1}\boldsymbol{\Omega}_t^{(1)}
\end{aligned} \tag{136}$$

$(\boldsymbol{\Omega}_t^{(1)})^\top(\mathbf{R}_{t,11})^{-1}\boldsymbol{\Omega}_t^{(1)}$ 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: the expected value of $\boldsymbol{Y}_t\boldsymbol{Y}_t^\top$

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

$$\begin{aligned}
\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}(1) = \boldsymbol{y}(1)] &= \mathrm{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}_t = \boldsymbol{y}_t] \\
&= \boldsymbol{y}_t\boldsymbol{y}_t^\top
\end{aligned} \tag{137}$$

If there are no observed values at time t, then

$$\begin{aligned}
&\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top] \\
&\quad= \mathrm{var}[\mathbf{Z}\boldsymbol{X}_t + \mathbf{a} + \mathbf{V}_t] + \mathrm{E}[\mathbf{Z}\boldsymbol{X}_t + \mathbf{a} + \mathbf{V}_t]\,\mathrm{E}[\mathbf{Z}\boldsymbol{X}_t + \mathbf{a} + \mathbf{V}_t]^\top \\
&\quad= \mathrm{var}[\mathbf{V}_t] + \mathrm{var}[\mathbf{Z}\boldsymbol{X}_t] + (\,\mathrm{E}[\mathbf{Z}\boldsymbol{X}_t + \mathbf{a}] + \mathrm{E}[\mathbf{V}_t])(\,\mathrm{E}[\mathbf{Z}\boldsymbol{X}_t + \mathbf{a}] + \mathrm{E}[\mathbf{V}_t])^\top \\
&\quad= \mathbf{R} + \mathbf{Z}\widetilde{\mathbf{V}}_t\mathbf{Z}^\top + (\mathbf{Z}\widetilde{\mathbf{x}}_t + \mathbf{a})(\mathbf{Z}\widetilde{\mathbf{x}}_t + \mathbf{a})^\top
\end{aligned} \tag{138}$$

When only some of the $\boldsymbol{Y}_t$ are observed, we use again the conditional probability of a multivariate normal (equation 132). From this property, we know

that

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

$$\begin{aligned}
\text{Thus }\ &\mathrm{var}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]\\
&= (\boldsymbol{\Omega}_t^{(2)})^\top(\mathbf{R}_{t,22}-\mathbf{R}_{t,21}(\mathbf{R}_{t,11})^{-1}\mathbf{R}_{t,12})\boldsymbol{\Omega}_t^{(2)}\\
&= (\boldsymbol{\Omega}_t^{(2)})^\top(\boldsymbol{\Omega}_t^{(2)}\mathbf{R}(\boldsymbol{\Omega}_t^{(2)})^\top - \boldsymbol{\Omega}_t^{(2)}\mathbf{R}(\boldsymbol{\Omega}_t^{(1)})^\top(\mathbf{R}_{t,11})^{-1}\boldsymbol{\Omega}_t^{(1)}\mathbf{R}(\boldsymbol{\Omega}_t^{(2)})^\top)\boldsymbol{\Omega}_t^{(2)}\\
&= \mathbf{I}_t^{(2)}(\mathbf{R}-\mathbf{R}(\boldsymbol{\Omega}_t^{(1)})^\top(\mathbf{R}_{t,11})^{-1}\boldsymbol{\Omega}_t^{(1)}\mathbf{R})\mathbf{I}_t^{(2)}\\
&= \mathbf{I}_t^{(2)}\nabla_t\mathbf{R}\mathbf{I}_t^{(2)}
\end{aligned} \tag{139}$$

The $\mathbf{I}_t^{(2)}$ bracketing both sides are 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.

$$\begin{aligned}
\widetilde{\mathbf{O}}_t =\ & \mathrm{E}_{XY}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)] = \mathrm{E}_X[\,\mathrm{E}_{Y|x}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]]\\
=\ & \mathrm{E}_X\big[\,\mathrm{var}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]\\
&+\ \mathrm{E}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]\,\mathrm{E}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]^\top\big]\\
=\ & \mathrm{E}_X[\mathbf{I}_t^{(2)}\nabla_t\mathbf{R}\mathbf{I}_t^{(2)}] + \mathrm{E}_X[(\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a}))(\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a}))^\top]\\
=\ & \mathbf{I}_t^{(2)}\nabla_t\mathbf{R}\mathbf{I}_t^{(2)} + \mathrm{var}_X\big[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a})\big]\\
&+\ \mathrm{E}_X[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a})]\,\mathrm{E}_X[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a})]^\top\\
=\ & \mathbf{I}_t^{(2)}\nabla_t\mathbf{R}\mathbf{I}_t^{(2)} + \mathbf{I}_t^{(2)}\nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_t\mathbf{Z}^\top\nabla_t^\top\mathbf{I}_t^{(2)} + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{y}}_t^\top
\end{aligned} \tag{140}$$

Thus,

$$\widetilde{\mathbf{O}}_t = \mathbf{I}_t^{(2)}(\nabla_t\mathbf{R}+\nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_t\mathbf{Z}^\top\nabla_t^\top)\mathbf{I}_t^{(2)} + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{y}}_t^\top \tag{141}$$

5.5 Derivation: 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

$$\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)] = \boldsymbol{y}_t\,\mathrm{E}[\boldsymbol{X}_t^\top] = \boldsymbol{y}_t\widetilde{\mathbf{x}}_t^\top \tag{142}$$

If there are no observed values at time t, then

$$\begin{aligned}
&\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)] \\
&\quad= \mathrm{E}[(\mathbf{Z}\boldsymbol{X}_t+\mathbf{a}+\mathbf{V}_t)\boldsymbol{X}_t^\top] \\
&\quad= \mathrm{E}[\mathbf{Z}\boldsymbol{X}_t\boldsymbol{X}_t^\top+\mathbf{a}\boldsymbol{X}_t^\top+\mathbf{V}_t\boldsymbol{X}_t^\top] \\
&\quad= \mathbf{Z}\widetilde{\mathbf{P}}_t+\mathbf{a}\widetilde{\mathbf{x}}_t^\top+\operatorname{cov}[\mathbf{V}_t,\boldsymbol{X}_t]+\mathrm{E}[\mathbf{V}_t]\,\mathrm{E}[\boldsymbol{X}_t]^\top \\
&\quad= \mathbf{Z}\widetilde{\mathbf{P}}_t+\mathbf{a}\widetilde{\mathbf{x}}_t^\top
\end{aligned} \tag{143}$$

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

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

$$\begin{aligned}
\widetilde{\mathbf{y}\mathbf{x}}_t &= \mathrm{E}_{XY}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)] \\
&= \operatorname{cov}[\boldsymbol{Y}_t,\boldsymbol{X}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1)]+\mathrm{E}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]\,\mathrm{E}_{XY}[\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]^\top \\
&= \operatorname{cov}[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a})+\mathbf{V}_t^*,\boldsymbol{X}_t]+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_t^\top \\
&= \operatorname{cov}[\boldsymbol{y}_t,\boldsymbol{X}_t]-\operatorname{cov}[\nabla_t\boldsymbol{y}_t,\boldsymbol{X}_t]+\operatorname{cov}[\nabla_t\mathbf{Z}\boldsymbol{X}_t,\boldsymbol{X}_t]+\operatorname{cov}[\nabla_t\mathbf{a},\boldsymbol{X}_t] \\
&\quad+\operatorname{cov}[\mathbf{V}_t^*,\boldsymbol{X}_t]+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_t^\top \\
&= 0-0+\nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_t+0+0+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_t^\top \\
&= \nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_t+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_t^\top
\end{aligned} \tag{144}$$

This uses the computational formula for covariance: $\mathrm{E}[\boldsymbol{Y}\boldsymbol{X}^\top]=\operatorname{cov}[\boldsymbol{Y},\boldsymbol{X}]+\mathrm{E}[\boldsymbol{Y}]\,\mathrm{E}[\boldsymbol{X}]^\top$. $\mathbf{V}_t^*$ 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 (135). $\mathbf{V}_t^*$ and $\boldsymbol{X}_t$ are independent of each other, thus $\operatorname{cov}[\mathbf{V}_t^*,\boldsymbol{X}_t^\top]=0$.

5.6 Derivation: 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 $\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_{t-1}^\top]$:

$$\begin{aligned}
\widetilde{\mathbf{y}\mathbf{x}}_t &= \mathrm{E}_{XY}[\boldsymbol{Y}_t\boldsymbol{X}_{t-1}^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)] \\
&= \operatorname{cov}[\boldsymbol{Y}_t,\boldsymbol{X}_{t-1}|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1)]+\mathrm{E}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]\,\mathrm{E}_{XY}[\boldsymbol{X}_{t-1}^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]^\top \\
&= \operatorname{cov}[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}\boldsymbol{X}_t-\mathbf{a})+\mathbf{V}_t^*,\boldsymbol{X}_{t-1}]+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_{t-1}^\top \\
&= \operatorname{cov}[\boldsymbol{y}_t,\boldsymbol{X}_{t-1}]-\operatorname{cov}[\nabla_t\boldsymbol{y}_t,\boldsymbol{X}_{t-1}]+\operatorname{cov}[\nabla_t\mathbf{Z}\boldsymbol{X}_t,\boldsymbol{X}_{t-1}] \\
&\quad+\operatorname{cov}[\nabla_t\mathbf{a},\boldsymbol{X}_{t-1}]+\operatorname{cov}[\mathbf{V}_t^*,\boldsymbol{X}_{t-1}]+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_{t-1}^\top \\
&= 0-0+\nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_{t,t-1}+0+0+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_{t-1}^\top \\
&= \nabla_t\mathbf{Z}\widetilde{\mathbf{V}}_{t,t-1}+\widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_{t-1}^\top
\end{aligned} \tag{145}$$

6 Degenerate variance modifications

It is possible that the model has deterministic and probabilistic elements; mathematically this means that one or the other of $\mathbf{R}$ or $\mathbf{Q}$ have zeros on the diagonal in which case some of the observation or state processes are deterministic. Assuming the model is solveable (one solution and not over-determined), we can modify the Kalman smoother and EM algorithm to handle models with deterministic elements.

As an example of a solveable versus unsolveable model, consider the following. If

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & a \neq 0 & 0 & 0 \\ 0 & 0 & b \neq 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{146}$$

then following are bad versus ok $\mathbf{Z}$ matrices.

$$\mathbf{Z}_{\text{bad}} = \begin{bmatrix} c & d & 0 \\ z(2,1) & z(2,2) & z(2,3) \\ z(3,1) & z(3,1) & z(3,1) \\ c & d & 0 \end{bmatrix}, \quad \mathbf{Z}_{\text{ok}} = \begin{bmatrix} c & 0 & 0 \\ z(2,1) & z(2,2) & z(2,3) \\ z(3,1) & z(3,1) & z(3,1) \\ c & d \neq 0 & 0 \end{bmatrix} \tag{147}$$

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

$$\begin{bmatrix} y_t(1) \\ y_t(4) \end{bmatrix} = \begin{bmatrix} cx_t(1) + dx_t(2) \\ cx_t(1) + dx_t(2) \end{bmatrix} \tag{148}$$

and since $y_t(1) \neq y_t(4)$, potentially, that is not solveable. The second $\mathbf{Z}$ reduces to

$$\begin{bmatrix} y_t(1) \\ y_t(4) \end{bmatrix} = \begin{bmatrix} cx_t(1) \\ cx_t(1) + dx_t(4) \end{bmatrix} \tag{149}$$

and that is solveable 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 $\mathbf{R}$ are equal to zero.

6.1 Kalman filter/smoother modifications

In principle, when one of the $\mathbf{Q}$ or $\mathbf{R}$ variances is zero[13], 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 will involve a matrix with a zero on the diagonal and this will lead to an error.

[13] The corresponding covariances will also be zero

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

$$\mathbf{K}_t = \mathbf{V}_t^{t-1}\mathbf{Z}_t^\top(\mathbf{Z}_t\mathbf{V}_t^{t-1}\mathbf{Z}_t^\top + \mathbf{R}_t)^{-1} \tag{150}$$

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 124). 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 $\mathbf{V}_1^0$ is 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 ($\boldsymbol{\Omega}_t^{(1)}$). Replace $\left(\mathbf{Z}_t\mathbf{V}_t^{t-1}\mathbf{Z}_t^\top + \mathbf{R}_t\right)^{-1}$ in equation (150) with

$$(\boldsymbol{\Omega}_t^{(1)})^\top\left(\boldsymbol{\Omega}_t^{(1)}(\mathbf{Z}_t\mathbf{V}_t^{t-1}\mathbf{Z}_t^\top + \mathbf{R}_t)(\boldsymbol{\Omega}_t^{(1)})^\top\right)^{-1}\boldsymbol{\Omega}_t^{(1)} \tag{151}$$

Wrapping in $\boldsymbol{\Omega}_t^{(1)}(\boldsymbol{\Omega}_t^{(1)})^\top$ 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 $(\boldsymbol{\Omega}_t^{(1)})^\top\boldsymbol{\Omega}_t^{(1)}$. 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 150 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 is

$$\begin{aligned}\mathbf{J}_t &= \mathbf{V}_{t-1}^{t-1}\mathbf{B}^\top(\mathbf{V}_t^{t-1})^{-1} \\ &\text{where } \mathbf{V}_t^{t-1} = \mathbf{B}\mathbf{V}_{t-1}^{t-1}\mathbf{B}^\top + \mathbf{Q}\end{aligned} \tag{152}$$

If $\mathbf{Q}$ has zeros on the diagonal, the corresponding $\mathbf{J}_t$ elements become:

$$\begin{aligned}\mathbf{J}_t &= \mathbf{V}_{t-1}^{t-1}\mathbf{B}^\top(\mathbf{B}\mathbf{V}_{t-1}^{t-1}\mathbf{B}^\top)^{-1} \\ &= \mathbf{V}_{t-1}^{t-1}\mathbf{B}^\top(\mathbf{B}^\top)^{-1}(\mathbf{V}_{t-1}^{t-1})^{-1}\mathbf{B}^{-1} \\ &= \mathbf{B}^{-1}\end{aligned} \tag{153}$$

Note when $\mathbf{V}_0 = 0$ and $\mathbf{Q} = 0$, $\mathbf{J}_0$ could be set to 0 or $\mathbf{B}^{-1}$.

6.2 EM algorithm modifications

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-invertable 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 129).

6.2.1 Modified likelihood for partially deterministic models

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

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & .2 \\ 0 & 0 & 0 \\ .2 & 0 & 1 \end{bmatrix}, \quad \text{then} \quad \mathbf{R}^+ = \begin{bmatrix} 1 & .2 \\ .2 & 1 \end{bmatrix}. \tag{154}$$

Let $\mathbf{\Omega}_r^+$ 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 $(\mathbf{\Omega}_r^+)^\top\mathbf{\Omega}_r^+ \equiv \mathbf{I}_r^+$ zero's out the zero row in $\mathbf{R}$ (and any $n \times 1$ row vector. For the example above,

$$\begin{aligned}\mathbf{R}^+ &= \mathbf{\Omega}_r^+\mathbf{R}(\mathbf{\Omega}_r^+)^\top \\ \boldsymbol{y}_t^+ &= \mathbf{\Omega}_r^+\boldsymbol{y}_t \\ \mathbf{\Omega}_r^+ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{I}_r^+ = (\mathbf{\Omega}_r^+)^\top\mathbf{\Omega}_r^+ = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}\end{aligned} \tag{155}$$

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

$$\begin{aligned}\mathbf{R}^{(0)} &= \mathbf{\Omega}_r^{(0)}\mathbf{R}(\mathbf{\Omega}_r^{(0)})^\top \\ \boldsymbol{y}_t^{(0)} &= \mathbf{\Omega}_r^{(0)}\boldsymbol{y}_t \\ \mathbf{\Omega}_r^{(0)} &= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \qquad \mathbf{I}_r^{(0)} = (\mathbf{\Omega}_r^{(0)})^\top\mathbf{\Omega}_r^{(0)} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\end{aligned} \tag{156}$$

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

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

$$\begin{aligned}
\boldsymbol{x}_t^{(0)} &= \mathbf{\Omega}_q^{(0)}\boldsymbol{x}_t = \mathbf{\Omega}_q^{(0)}\mathbf{B}\boldsymbol{x}_{t-1} + \mathbf{\Omega}_q^{(0)}\mathbf{u} \\
\boldsymbol{x}_t^+ &= \mathbf{\Omega}_q^+\boldsymbol{x}_t = \mathbf{\Omega}_q^+\mathbf{B}\boldsymbol{x}_{t-1} + \mathbf{\Omega}_q^+\mathbf{u} + \mathbf{w}_t^+ \\
\mathbf{w}_t^+ &\sim \text{MVN}(0, \mathbf{Q}^+) \\
\boldsymbol{x}_0 &\sim \text{MVN}(\boldsymbol{\xi}, \mathbf{V}_0) \\
\boldsymbol{y}_t^{(0)} &= \mathbf{\Omega}_r^{(0)}\boldsymbol{y}_t = \mathbf{\Omega}_r^{(0)}(\mathbf{Z}\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{Z}(\mathbf{\Omega}_q^{(0)})^\top\mathbf{\Omega}_q^{(0)}(\mathbf{B}\boldsymbol{x}_{t-1} + \mathbf{u}) + \mathbf{a}) \\
\boldsymbol{y}_t^+ &= \mathbf{\Omega}_q^+\boldsymbol{y}_t = \mathbf{\Omega}_q^+(\mathbf{Z}\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{Z}\mathbf{I}_q^{(0)}(\mathbf{B}\boldsymbol{x}_{t-1} + \mathbf{u}) + \mathbf{a}) + \mathbf{v}_t^+ \\
\mathbf{v}_t^+ &\sim \text{MVN}(0, \mathbf{R}^+)
\end{aligned} \tag{157}$$

In order for this to be solveable, we require that $\mathbf{\Omega}_r^{(0)}\mathbf{Z}\mathbf{I}_q^{(0)}$ is all zeros so that $\mathbf{B}$ and $\mathbf{u}$ do not appear in the $\boldsymbol{y}^{(0)}$ equation, and then they disappear in the $\boldsymbol{y}^{(0)}$ equation as shown above. That is, if state process x_i (in $\boldsymbol{x}$) is deterministic (0 process variance), then no observation processes y in $\boldsymbol{y}$ that involves that x_i (through $\mathbf{Z}$) shall have 0 observation variance. Also notice that $\mathbf{\Omega}_r^{(0)}\mathbf{Z}$ and $\mathbf{\Omega}_r^{(0)}\mathbf{a}$ appear in the $\boldsymbol{y}^{(0)}$ equation and not in the $\boldsymbol{y}^+$ equation. This means that $\mathbf{\Omega}_r^{(0)}\mathbf{Z}$ and $\mathbf{\Omega}_r^{(0)}\mathbf{a}$ cannot be estimated but must be fixed terms.

Summarizing, this equation becomes

$$\begin{aligned}
\boldsymbol{x}_t^{(0)} &= \mathbf{B}^{(0)}\boldsymbol{x}_{t-1} + \mathbf{u}^{(0)} \\
\boldsymbol{x}_t^+ &= \mathbf{B}^+\boldsymbol{x}_{t-1} + \mathbf{u}^+ + \mathbf{w}_t^+ \\
\mathbf{w}_t^+ &\sim \text{MVN}(0, \mathbf{Q}^+) \\
\boldsymbol{x}_0 &\sim \text{MVN}(\boldsymbol{\xi}, \mathbf{V}_0) \\
\boldsymbol{y}_t^{(0)} &= \mathbf{Z}^{(0)}\boldsymbol{x}_t + \mathbf{a}^{(0)} \\
\boldsymbol{y}_t^+ &= \mathbf{Z}^+\boldsymbol{x}_t + \mathbf{a}^+ + \mathbf{v}_t^+ \\
&= \mathbf{Z}^+\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{Z}^+\mathbf{I}_q^{(0)}\boldsymbol{x}_t + \mathbf{a}^+ + \mathbf{v}_t^+ \\
\mathbf{v}_t^+ &\sim \text{MVN}(0, \mathbf{R}^+)
\end{aligned} \tag{158}$$

As discussed above, we require that $\mathbf{\Omega}_r^{(0)}\mathbf{Z}\mathbf{I}_q^{(0)}$ is all zeros while $\mathbf{\Omega}_r^+\mathbf{Z}\mathbf{I}_q^{(0)}$ has no rows that are all zeros. This equation is conceptually the same as equation 4.2.28 in Harvey (1989).

We want to write down the joint likelihood of $\boldsymbol{y}^+ = \{\boldsymbol{y}_1^+, \boldsymbol{y}_2+, \boldsymbol{y}_3+, ...\}$ and $\boldsymbol{x}^+ = \{\boldsymbol{x}_1^+, \boldsymbol{x}_2^+, \boldsymbol{x}_3^+, ...\}$. We can write the joint log-likelihood function for the + elements using equation (157) and the likelihood function for a multivariate

normal distribution.

$$
\begin{aligned}
&\Psi^+ = \log \mathbf{L}(\boldsymbol{y}^+, \boldsymbol{x}^+|\Theta) = \\
&-\frac{1}{2}\sum_1^T (\boldsymbol{y}_t^+ - \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{I}_q^{(0)}\boldsymbol{x}_t) - \mathbf{a}^+)^\top (\mathbf{R}^+)^{-1} \\
&\quad (\boldsymbol{y}_t^+ - \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{I}_q^{(0)}\boldsymbol{x}_t) - \mathbf{a}^+) - \frac{T}{2}\log|\mathbf{R}^+| \\
&-\frac{1}{2}\sum_1^T (\boldsymbol{x}_t^+ - \mathbf{B}^+\boldsymbol{x}_{t-1} - \mathbf{u}^+)^\top (\mathbf{Q}^+)^{-1}(\boldsymbol{x}_t^+ - \mathbf{B}^+\boldsymbol{x}_{t-1} - \mathbf{u}^+) - \frac{T}{2}\log|\mathbf{Q}^+| \\
&-\frac{1}{2}(\boldsymbol{x}_0 - \boldsymbol{\xi})^\top \mathbf{V}_0^{-1}(\boldsymbol{x}_0 - \boldsymbol{\xi}) - \frac{1}{2}\log|\mathbf{V}_0| - \frac{n}{2}\log 2\pi
\end{aligned}
\tag{159}
$$

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 while $\mathbf{B}^+$ and $\mathbf{u}^+$ appear in the $\boldsymbol{x}$ part.

If $\boldsymbol{x}_0$ is treated as fixed ($\mathbf{V}_0 = 0$), then the likelihood takes a slightly different form using equation (158)

$$
\begin{aligned}
&\Psi^+ = \log \mathbf{L}(\boldsymbol{y}^+, \boldsymbol{x}^+|\Theta) = \\
&-\frac{1}{2}\sum_1^T (\boldsymbol{y}_t^+ - \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{I}_q^{(0)}\boldsymbol{x}_t) - \mathbf{a}^+)^\top (\mathbf{R}^+)^{-1} \\
&\quad (\boldsymbol{y}_t^+ - (\mathbf{Z}^+\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{I}_q^{(0)}\boldsymbol{x}_t) - \mathbf{a}^+) - \frac{T}{2}\log|\mathbf{R}^+| \\
&-\frac{1}{2}\sum_1^T (\boldsymbol{x}_t^+ - \mathbf{B}^+\boldsymbol{x}_{t-1} - \mathbf{u}^+)^\top (\mathbf{Q}^+)^{-1}(\boldsymbol{x}_t^+ - \mathbf{B}^+\boldsymbol{x}_{t-1} - \mathbf{u}^+) \\
&-\frac{T}{2}\log|\mathbf{Q}^+| - \frac{n}{2}\log 2\pi \\
&\quad \text{where } \boldsymbol{x}_0 \equiv \boldsymbol{\xi}
\end{aligned}
\tag{160}
$$

6.2.2 $\mathbf{Z}^+$ 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 $\mathbf{Z}$ (108) with $\mathbf{R}^{-1}$ replaced with $\mathbf{R}^*$:

$$\begin{aligned}\boldsymbol{\zeta}_{j+1} = &\left(\sum_{t=1}^{T}(\mathbf{D}_z^\top(\widetilde{\mathbf{P}}_t \otimes \mathbf{R}^*)\mathbf{D}_z)\right)^{-1}\mathbf{D}_z^\top \times \\ &\sum_{t=1}^{T}\left(\operatorname{vec}(\mathbf{R}^*(\widetilde{\mathbf{y}\mathbf{x}}_t - \mathbf{a}\widetilde{\mathbf{x}}_t^\top)) - (\widetilde{\mathbf{P}}_t \otimes \mathbf{R}^*)\mathbf{f}_z\right)\end{aligned} \tag{161}$$

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

$$\operatorname{vec}(\mathbf{Z}_{j+1}) = \mathbf{f}_z + \mathbf{D}_z\boldsymbol{\zeta}_{j+1} \tag{162}$$

Because the $\mathbf{Z}^{(0)}$ elements are fixed, $\mathbf{D}_z^\top(\widetilde{\mathbf{P}}_t \otimes \mathbf{R}^*)\mathbf{D}_z$ is invertable. 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 86) with $\mathbf{R}^{-1}$ replaced with $\mathbf{R}^*$:

$$\boldsymbol{\alpha}_{j+1} = \frac{1}{T}\left(\mathbf{D}_a^\top\mathbf{R}^*\mathbf{D}_a\right)^{-1}\mathbf{D}_a^\top\mathbf{R}^*\sum_{t=1}^{T}\left(\widetilde{\mathbf{y}}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{f}_a\right) \tag{163}$$

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

$$\mathbf{a}_{j+1} = \mathbf{f}_a + \mathbf{D}_a\boldsymbol{\alpha}_{j+1}, \tag{164}$$

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

6.2.3 u update equation for partially deterministic models with diagonal $\mathbf{B}^{(0)}$

To derive the update equation for $\mathbf{u}$, we need to take the partial derivative of Ψ^+ holding everything constant except $\mathbf{u}$. If a state process is fully deterministic and $\mathbf{B}$ is diagonal, then we cannot hold $\boldsymbol{x}^{(0)}$ constant while changing $\mathbf{u}^{(0)}$. If we change $\mathbf{u}^{(0)}$, then $\boldsymbol{x}^{(0)}$ must change because it is deterministic. This is in contrast to $\mathbf{u}^+$ which can be changed while holding $\boldsymbol{x}^+$ constant, because $\boldsymbol{x}^+$ is stochastic and all values are possible for a given $\mathbf{u}^+$ (albeit maybe not as likely).

When $\mathbf{B}^{(0)}$ is diagonal and there are no $\mathbf{B}$ elements linking the $\mathbf{B}^+$ and $\mathbf{B}^{(0)}$ blocks, the equation for the deterministic state processes becomes $x_t = bx_{t-1}+u$ which is simply

$$\begin{aligned}x_t = &b^t x_0 + u\sum_{i=0}^{t-1} b^i = \\ &b^t x_0 + u\frac{1-b^t}{1-b}, b \neq 1 \\ &x_0 + ut, b = 1\end{aligned} \tag{165}$$

Thus we will replace the $\mathbf{I}_q^{(0)}\boldsymbol{x}_t$ term appearing in Ψ^+ (equation 159) with

$$\begin{aligned}\mathbf{I}_q^{(0)}\boldsymbol{x}_t &= (\mathbf{B}^{(0)})^t\boldsymbol{x}_0^{(0)} + (\mathbf{I}_q^{(0)} - (\mathbf{B}^{(0)})^t)(\mathbf{I}-\mathbf{B}^{(0)})^{-1}\mathbf{u}^{(0)} \\ &= \mathbf{B}^{\diamondsuit}\boldsymbol{x}_0 + \mathbf{B}^{\sharp}\mathbf{u} \\ &\text{where } \mathbf{B}^{\diamondsuit} = (\mathbf{I}_q^{(0)}\mathbf{B}\mathbf{I}_q^{(0)})^t \\ &\text{and } \mathbf{B}^{\sharp} = (\mathbf{I}_q^{(0)} - \mathbf{B}^{\diamondsuit})(\mathbf{I}_m - \mathbf{I}_q^{(0)}\mathbf{B}\mathbf{I}_q^{(0)})^{-1}\end{aligned} \tag{166}$$

Thus Ψ^+ becomes

$$\begin{aligned}&\Psi^+ = \log \mathbf{L}(\boldsymbol{y}^+, \boldsymbol{x}^+|\Theta) = \\ &-\frac{1}{2}\sum_1^T(\boldsymbol{y}_t^+ - \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{I}_q^{(0)}(\mathbf{B}^{\diamondsuit}\boldsymbol{x}_0 + \mathbf{B}^{\sharp}\mathbf{u})) - \mathbf{a}^+)^\top(\mathbf{R}^+)^{-1} \\ &\quad(\boldsymbol{y}_t^+ - \mathbf{Z}^+(\mathbf{I}_q^+\boldsymbol{x}_t + \mathbf{I}_q^{(0)}(\mathbf{B}^{\diamondsuit}\boldsymbol{x}_0 + \mathbf{B}^{\sharp}\mathbf{u})) - \mathbf{a}^+) - \frac{T}{2}\log|\mathbf{R}^+| \\ &-\frac{1}{2}\sum_1^T(\boldsymbol{x}_t^+ - \mathbf{B}^+\boldsymbol{x}_{t-1} - \mathbf{u}^+)^\top(\mathbf{Q}^+)^{-1}(\boldsymbol{x}_t^+ - \mathbf{B}^+\boldsymbol{x}_{t-1} - \mathbf{u}^+) \\ &-\frac{T}{2}\log|\mathbf{Q}^+| - \frac{1}{2}(\boldsymbol{x}_0 - \boldsymbol{\xi})^\top\mathbf{V}_0^{-1}(\boldsymbol{x}_0 - \boldsymbol{\xi}) - \frac{1}{2}\log|\mathbf{V}_0| - \frac{n}{2}\log 2\pi\end{aligned} \tag{167}$$

This works because $\mathbf{I}_q^{(0)}\mathbf{B}\mathbf{I}_q^{(0)}$ is diagonal. The $\mathbf{u}^{(0)}$ parameter appears in the $\boldsymbol{y}$ part of the likelihood 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 then $\mathbf{u}$ as $\mathbf{f}_u + \mathbf{D}_u\boldsymbol{v}$, put that in equation (167), and differentiate with respect to $\boldsymbol{v}$ rather than $\mathbf{u}^{(0)}$ or $\mathbf{u}^+$.

The rest of the derivation steps are similar to those for the general update equation (analogous to equation 84). Take the derivative of Ψ^+ (equation 167) with respect to $\boldsymbol{v}$. Note that $\mathbf{I}_q^{(0)} = (\mathbf{I}_q^{(0)})^\top$ and that $\mathbf{I}_q^{(0)}(\mathbf{I}-\mathbf{B}^t)(\mathbf{I}-\mathbf{B})^{-1}\mathbf{I}_q^{(0)}$ is a diagonal matrix and thus can be moved (i.e., if A and D are diagonal, $AD = DA$ and $D^\top = D$). After taking the derivative with respect to $\boldsymbol{v}$, we get:

$$\begin{aligned}&\mathbf{D}_u^\top(\mathbf{R}^\sharp + T\mathbf{Q}^*)\mathbf{D}_u\boldsymbol{v} = \\ &\quad\mathbf{D}_u^\top\mathbf{I}_q^{(0)}\sum_{t=1}^T\mathbf{B}^\sharp\mathbf{Z}^\top\mathbf{R}^*\big(\widetilde{\mathbf{y}}_t - \mathbf{Z}\mathbf{I}_q^+\widetilde{\mathbf{x}}_t - \mathbf{Z}\mathbf{I}_q^{(0)}(\mathbf{B}^{\diamondsuit}\widetilde{\mathbf{x}}_0 + \mathbf{B}^\sharp\mathbf{f}_u) - \mathbf{a}\big) \\ &\quad+\mathbf{D}_u^\top\mathbf{I}_q^+\mathbf{Q}^*\sum_{t=1}^T\big(\widetilde{\mathbf{x}}_t - \mathbf{B}\widetilde{\mathbf{x}}_{t-1} - \mathbf{f}_u\big) \\ &\text{where } \mathbf{R}^* = (\boldsymbol{\Omega}_r^+)^\top(\mathbf{R}^+)^{-1}\boldsymbol{\Omega}_r^+ \\ &\text{and } \mathbf{R}^\sharp = \sum_{t=1}^T\mathbf{B}^\sharp\mathbf{Z}^\top\mathbf{R}^*\mathbf{Z}\mathbf{B}^\sharp \\ &\text{and } \mathbf{Q}^* = (\boldsymbol{\Omega}_q^+)^\top(\mathbf{Q}^+)^{-1}\boldsymbol{\Omega}_q^+\end{aligned} \tag{168}$$

Again, this update equation is based on constraining $\mathbf{I}_q^{(0)}\mathbf{B}\mathbf{I}_q^{(0)}$ to be diagonal and $\mathbf{I}_q^{+}\mathbf{B}\mathbf{I}_q^{(0)}$ and $\mathbf{I}_q^{(0)}\mathbf{B}\mathbf{I}_q^{+}$ to be zero. This means that $\mathbf{B}$ can be rearranged to look like so, where the c's show the $\mathbf{B}^{(0)}$ block and the b's show the $\mathbf{B}^{+}$ block:

$$\begin{bmatrix} c & 0 & 0 & 0 & 0 \\ 0 & c & 0 & 0 & 0 \\ 0 & 0 & b & b & b \\ 0 & 0 & b & b & b \\ 0 & 0 & b & b & b \end{bmatrix} \tag{169}$$

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 and the corresponding row/column of $\mathbf{Z}^{\top}\mathbf{R}^{*}\mathbf{Z}$ will be non-zero. 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 $\boldsymbol{v}$ is

$$\begin{aligned} \boldsymbol{v}_{j+1} = &\left(\mathbf{D}_u^{\top}(\mathbf{R}^{\sharp} + T\mathbf{Q}^{*})\mathbf{D}_u\right)^{-1}\mathbf{D}_u^{\top}\times \\ &\Bigg(\sum_{t=1}^{T}\mathbf{B}^{\sharp}\mathbf{Z}^{\top}\mathbf{R}^{*}\left(\widetilde{\mathbf{y}}_t - \mathbf{Z}\mathbf{I}_q^{+}\widetilde{\mathbf{x}}_t - \mathbf{Z}\mathbf{B}^{\diamondsuit}\widetilde{\mathbf{x}}_0 - \mathbf{Z}\mathbf{B}^{\sharp}\mathbf{f}_u - \mathbf{a}\right) \\ &+ \mathbf{I}_q^{+}\mathbf{Q}^{*}\sum_{t=1}^{T}\left(\widetilde{\mathbf{x}}_t - \mathbf{B}\widetilde{\mathbf{x}}_{t-1} - \mathbf{f}_u\right)\Bigg) \end{aligned} \tag{170}$$

and

$$\mathbf{u}_{j+1} = \mathbf{f}_u + \mathbf{D}_u\boldsymbol{v}_{j+1}, \tag{171}$$

where $\mathbf{B}^{\diamondsuit}$ and $\mathbf{B}^{\sharp}$ are defined in equation (166) and $\mathbf{R}^{\sharp}$, $\mathbf{R}^{*}$ and $\mathbf{Q}^{*}$ are defined in equation (168). If $\boldsymbol{x}_0$ is treated as fixed, $\widetilde{\mathbf{x}}_0$ is replaced with $\boldsymbol{\xi}$, otherwise it has its usual definition ($\mathrm{E}[\boldsymbol{x}_0|\boldsymbol{y}, \Theta]$).

Conceptually, the approach described here is the same as the approach presented in section 4.2.5 of (Harvey, 1989), but it is a little more general because it deals with the case where some $\mathbf{u}$ elements are shared, 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 $j + 1$.

6.2.4 $\boldsymbol{\xi}$ update equation for partially deterministic models with diagonal $\mathbf{B}^{(0)}$

Take the derivative of Ψ^+ (equation 167) with respect to $\mathbf{p}$ where $\boldsymbol{\xi} = \mathbf{f}_\xi + \mathbf{D}_\xi \mathbf{p}$. The constrained $\mathbf{p}$ update equation when $\mathbf{Q}$ has zeros on the diagonal is then

$$\begin{aligned}
&\mathbf{D}_\xi^\top (\mathbf{R}^\diamondsuit + \mathbf{B}^\top \mathbf{Q}^* \mathbf{B}) \mathbf{D}_\xi \mathbf{p} = \\
&\quad \mathbf{D}_\xi^\top \Bigg(\mathbf{I}_q^{(0)} \sum_{t=1}^{T} \mathbf{B}^\diamondsuit \mathbf{Z}^\top \mathbf{R}^* \big(\widetilde{\mathbf{y}}_t - \mathbf{Z} \mathbf{I}_q^+ \widetilde{\mathbf{x}}_t - \mathbf{Z} \mathbf{B}^\sharp \mathbf{u} - \mathbf{Z} \mathbf{B}^\diamondsuit \mathbf{f}_\xi - \mathbf{a} \big) \\
&\quad + \mathbf{I}_q^+ \mathbf{B}^\top \mathbf{Q}^* \big(\widetilde{\mathbf{x}}_1 - \mathbf{B} \mathbf{f}_\xi - \mathbf{u} \big) \Bigg) \\
&\text{where } \mathbf{R}^\diamondsuit = \sum_{t=1}^{T} \mathbf{B}^\diamondsuit \mathbf{Z}^\top \mathbf{R}^* \mathbf{Z} \mathbf{B}^\diamondsuit
\end{aligned} \tag{172}$$

The matrices $\mathbf{B}^\diamondsuit$ and $\mathbf{B}^\sharp$ are defined in equation (166), and $\mathbf{R}^*$ and $\mathbf{Q}^*$ are defined in equation (168). $\mathbf{B}^{(0)}$ is constrained to be diagonal.

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

$$\begin{aligned}
\mathbf{p}_{j+1} = &\big(\mathbf{D}_\xi^\top (\mathbf{R}^\diamondsuit + \mathbf{B}^\top \mathbf{Q}^* \mathbf{B}) \mathbf{D}_\xi \big)^{-1} \mathbf{D}_\xi^\top \times \\
&\Bigg(\mathbf{I}_q^{(0)} \sum_{t=1}^{T} \mathbf{B}^\diamondsuit \mathbf{Z}^\top \mathbf{R}^* \big(\widetilde{\mathbf{y}}_t - \mathbf{Z} \mathbf{I}_q^+ \widetilde{\mathbf{x}}_t - \mathbf{Z} \mathbf{B}^\sharp \mathbf{u} - \mathbf{Z} \mathbf{B}^\diamondsuit \mathbf{f}_\xi - \mathbf{a} \big) \\
&\quad + \mathbf{I}_q^+ \mathbf{B}^\top \mathbf{Q}^* \big(\widetilde{\mathbf{x}}_1 - \mathbf{B} \mathbf{f}_\xi - \mathbf{u} \big) \Bigg)
\end{aligned} \tag{173}$$

and

$$\boldsymbol{\xi}_{j+1} = \mathbf{f}_\xi + \mathbf{D}_\xi \mathbf{p}_{j+1}, \tag{174}$$

6.2.5 B update equation for partially deterministic models when $\mathbf{B}^{(0)}$ is diagonal

First we would write Ψ^+ in equation (167) 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 and use relation (72) to show that:

$$\begin{aligned}
&\mathbf{I}_q^{(0)} \mathbf{B}^\diamondsuit \mathbf{I}_q^{(0)} \boldsymbol{X}_0 = (\boldsymbol{X}_0^\top \otimes \mathbf{I})((\mathbf{f}_b^{(0)})^t + \mathbf{D}_b^{(0)} \boldsymbol{\beta}^t), \\
&\mathbf{I}_q^{(0)} \mathbf{B}^\sharp \mathbf{I}_q^{(0)} \mathbf{u} = (\mathbf{u}^\top \otimes \mathbf{I})((\mathbf{f}_b^{(0)})^\sharp + \mathbf{D}_b^{(0)} \boldsymbol{\beta}^\sharp), \\
&\text{where } \mathbf{d}^t \equiv \begin{bmatrix} d_1^t \\ d_2^t \\ \cdots \\ d_p^t \end{bmatrix} \\
&\text{where } \mathbf{d}^\sharp \equiv \begin{bmatrix} d_1^t/(1-d_1) \\ d_2^t/(1-d_2) \\ \cdots \\ d_p^t/(1-d_p) \end{bmatrix}
\end{aligned} \tag{175}$$

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

The derivation proceeds by taking the derivative of Ψ^+ with respect to $\boldsymbol{\beta}$. However we will 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 solve that problem. 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[14] 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 $\boldsymbol{\xi}$ and $\mathbf{V}_1$ can be estimated though not simultaneously. Harvey (1989), 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, $\boldsymbol{\xi}$ is XX_1 and is then estimated as a parameter; $\mathbf{V}_1$ is held fixed at 0. In case 2, $\boldsymbol{\xi}$ nor $\mathbf{V}_1$ specify the mean and variance of $\boldsymbol{X}_1$. 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, 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 correct correlation structure. 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 really get around this since it has a correlation structure also.

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

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 apriori 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 apriori 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, auxilliary 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.