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

Elizabeth Eli Holmes*

original 16 Feb 2013
typo in Eqn 124, 14 Apr 2018
added derivation of variance of Y conditioned on y and x, 3 Feb 2020
typo in eqn 29-31, 20 Oct 2020

Abstract

This report presents an Expectation-Maximization (EM) algorithm for estimation of the maximum-likelihood parameter values of constrained multivariate autoregressive Gaussian state-space (MARSS) models. The MARSS model can be written: x(t)=Bx(t-1)+u+w(t), y(t)=Zx(t)+a+v(t), where w(t) and v(t) are multivariate normal error-terms with variance-covariance matrices Q and R respectively. MARSS models are a class of dynamic linear model and vector autoregressive model state-space model. Shumway and Stoffer presented an unconstrained EM algorithm for this class of models in 1982, and a number of researchers have presented EM algorithms for specific types of constrained MARSS models since then. In this report, I present a general EM algorithm for constrained MARSS models, where the constraints are on the elements within the parameter matrices (B,u,Q,Z,a,R). The constraints take the form vec(M)=f+Dm, where M is the parameter matrix, f is a column vector of fixed values, D is a matrix of multipliers, and m is the column vector of estimated values. This allows a wide variety of constrained parameter matrix forms. The presentation is for a time-varying MARSS model, where time-variation enters through the fixed (meaning not estimated) f(t) and D(t) matrices for each parameter. The algorithm allows missing values in y and partially deterministic systems where 0s appear on the diagonals of Q or R.

Keywords: Time-series analysis, Kalman filter, EM algorithm, maximum-likelihood, vector autoregressive model, dynamic linear model, parameter estimation, state-space

citation: Holmes, E. E. 2013. Derivation of the EM algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models. Technical Report. arXiv:1302.3919

*Northwest Fisheries Science Center, NOAA Fisheries, Seattle, WA 98112, eli.holmes@noaa.gov, http://faculty.washington.edu/eeholmes

1 Overview

EM algorithms extend maximum-likelihood estimation to models with hidden states and are widely used in engineering and computer science applications. This report presents an EM algorithm for a general class of Gaussian constrained multivariate autoregressive state-space (MARSS) models, with a hidden multivariate autoregressive process (state) model and a multivariate observation model. This is an important class of time-series model used in many different scientific fields. 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.

Before showing the derivation for the constrained case, I first show a derivation of the EM algorithm for unconstrained[1] MARSS model. This EM algorithm was published by Shumway and Stoffer (1982), but my derivation is more similar to Ghahramani et al's (Ghahramani and Hinton, 1996; Roweis and Ghahramani, 1999) slightly different presentation. One difference in my presentation and all these previous presentations, however, 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. Another difference is that I present the update equations for both stochastic initial states and fixed initial states. I then extend the derivation to constrained MARSS models where there are fixed and shared elements in the parameter matrices and to the case of degenerate MARSS models where some processes in the model are deterministic rather than stochastic. 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.

When working with MARSS models, one should be cognizant that misspecification of the prior on the initial hidden states can have catastrophic and difficult to detect effects on the parameter estimates. There is often no sign that something is amiss with the MLE estimates output by an EM algorithm. There has been much work on how to avoid these initial conditions effects; see especially literature on vector autoregressive state-space models in the economics literature. The trouble often occurs when the prior on the initial states is inconsistent with the distribution of the initial states that is implied by the maximum-likelihood model. This often happens when the model implies a specific covariance structure on the initial states, but since the maximum-likelihood parameters are unknown, this covariance structure is unknown. Using a diffuse prior does not help since your diffuse prior still has some covariance structure (often independence is being imposed). In some ways the EM algorithm is less sensitive to a misspecified prior because it uses the smoothed states conditioned on all the data. However, if the prior is inconsistent with the model, the EM algorithm will not (cannot) find the MLEs. 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 derivation presented here allows one to circumvent these problems by treating the initial states as fixed (and estimated) parameters. The problematic initial state variance-covariance matrix is removed from the model, albeit at the cost of additional estimated parameters.

Finally, when working with MARSS models, one needs to ensure that the model is identifiable; i.e., a unique solution exists. For a given MARSS model, some of the parameter elements will need to be fixed (not estimated) in order to produce a model with one solution. How to do that depends on the MARSS model being fitted and is up to the user.

1.1 The MARSS model

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

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

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

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

The $\boldsymbol{y}$ equation is called the observation process, and $\boldsymbol{y}_t$ is a $n \times 1$ vector. The $\boldsymbol{x}$ equation is called the state or process equation, and $\boldsymbol{x}_t$ is a $m \times 1$ vector. The equation for $\boldsymbol{x}$ describes a multivariate autoregressive process (also called a random walk or Markov process). $\boldsymbol{w}$ are the process errors and are specific realizations of the random variable $\boldsymbol{W}$; $\boldsymbol{v}$ is defined similarly. The initial state can either defined at $t = 0$, as is done in

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

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

equation 1, or at $t = 1$. When presenting the MARSS model, I use $t = 0$ but the derivations will show the EM algorithm for both cases. $\mathbf{Q}$ and $\mathbf{R}$ are variance-covariance matrices that specify the stochasticity in the observation and state equations.

In the MARSS model, $\boldsymbol{x}$ and $\boldsymbol{y}$ equations describe two stochastic processes. By tradition, one conditions on observations of $\boldsymbol{y}$, and $\boldsymbol{x}$ is treated as completely hidden, hence the name 'hidden Markov process' of which a MARSS model is a special type. However, you could condition on (partial) observations of $\boldsymbol{x}$ and treat $\boldsymbol{y}$ as a (partially) hidden process—with as usual proper constraints to ensure identifiability. Nonetheless in this report, I follow tradition and treat $\boldsymbol{x}$ as hidden and $\boldsymbol{y}$ as (partially) observed. If $\boldsymbol{x}$ is partially observed then the update equations stay the same but the expectations shown in section 6 would be computed conditioned on the partially observed $\boldsymbol{x}$.

The first part of this report will review the derivation of an EM algorithm for the time-constant MARSS model (equation 1). However the main objective of this report is to show the derivation of an EM algorithm to solve a much more general MARSS model (section 4), which is a MARSS model with linear constraints on time-varying parameters:

$$\begin{aligned} \boldsymbol{x}_t &= \mathbf{B}_t\boldsymbol{x}_{t-1} + \mathbf{u}_t + \mathbf{G}_t\boldsymbol{w}_t, \text{ where } \boldsymbol{W}_t \sim \text{MVN}(0, \mathbf{Q}_t) \\ \boldsymbol{y}_t &= \mathbf{Z}_t\boldsymbol{x}_t + \mathbf{a}_t + \mathbf{H}_t\boldsymbol{v}_t, \text{ where } \boldsymbol{V}_t \sim \text{MVN}(0, \mathbf{R}_t) \\ \boldsymbol{x}_0 &= \boldsymbol{\xi} + \mathbf{F}\boldsymbol{l}, \text{ where } \boldsymbol{l} \sim \text{MVN}(0, \boldsymbol{\Lambda}) \end{aligned} \tag{2}$$

The initial state can either defined at $t = 0$, as is done in equation 2, or at $t = 1$.

The linear constraints appear as the vectorization of each parameter ($\mathbf{B}$, $\mathbf{u}$, $\mathbf{Q}$, $\mathbf{Z}$, $\mathbf{a}$, $\mathbf{R}$, $\boldsymbol{\xi}$, $\boldsymbol{\Lambda}$) is described by the relation $\mathbf{f}_t + \mathbf{D}_t\mathbf{m}$. This relation specifies linear constraints of the form $\beta_i + \beta_{a,i}a + \beta_{b,i}b + \dots$ on the elements in each MARSS parameter matrix. Equation 2 is a much broader class of MARSS models that includes MARSS models with exogenous variable (covariates), AR-p models, moving average models, constrained MARSS models and models that are combinations of these. The derivation also includes partially deterministic systems where $\mathbf{G}_t$, $\mathbf{H}_t$ and $\mathbf{F}$ may have all zero rows.

1.2 The joint log-likelihood function

Equation 2 describes a multivariate stochastic process and $\boldsymbol{Y}_t$ and $\boldsymbol{X}_t$ are random variables whose distributions are given by Equation 2. Denote a specific realization of these random variables as $\boldsymbol{y}$ and $\boldsymbol{x}$ which denotes a set of all y's and x's from $t = 1$ to T. The joint log-likelihood[3] of $\boldsymbol{y}$ and $\boldsymbol{x}$ can then be written then as follows[4], where $\boldsymbol{X}_t$ denotes the random variable and $\boldsymbol{x}_t$ is a realization from that random variable (and similarly for $\boldsymbol{Y}_t$):[5]

$$f(\boldsymbol{y}, \boldsymbol{x}) = f(\boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x})f(\boldsymbol{x}), \tag{3}$$

where

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

Thus,

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{x}) &= \prod_{t=1}^{T} f(\boldsymbol{y}_t|\boldsymbol{X} = \boldsymbol{x}) \times f(\boldsymbol{x}_0)\prod_{t=1}^{T} f(\boldsymbol{x}_t|\boldsymbol{X}_1^{t-1} = \boldsymbol{x}_1^{t-1}) \\ &= \prod_{t=1}^{T} f(\boldsymbol{y}_t|\boldsymbol{X}_t = \boldsymbol{x}_t) \times f(\boldsymbol{x}_0)\prod_{t=1}^{T} f(\boldsymbol{x}_t|\boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1}). \end{aligned} \tag{5}$$

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

[4] The log-likelihood function is shown here for the MARSS with non-time varying parameters (equation 1).

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

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 $\boldsymbol{v}_t$ 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, section 4.3).

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

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 $\boldsymbol{\Lambda}$. The likelihood then is slightly different. $\boldsymbol{x}_0$ is defined as a parameter $\boldsymbol{\xi}$ and

$$\begin{aligned}\log \boldsymbol{L}(\boldsymbol{y}, \boldsymbol{x}; \Theta) = &-\sum_1^T \frac{1}{2}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a})^\top \mathbf{R}^{-1}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a}) - \sum_1^T \frac{1}{2}\log|\mathbf{R}| \\ &-\sum_1^T \frac{1}{2}(\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u})^\top \mathbf{Q}^{-1}(\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u}) - \sum_1^T \frac{1}{2}\log|\mathbf{Q}|\end{aligned} \tag{7}$$

$\boldsymbol{\xi}$ appears in the likelihood for $\boldsymbol{x}_{t-1}$ when $t = 1$ in the summation. Note that in this case, $\boldsymbol{x}_0$ is no longer a realization of a random variable $\boldsymbol{X}_0$; it is a fixed (but unknown) parameter. Equation 7 is written as if all the $\boldsymbol{x}_0$ are fixed, however when the general derivation is presented, it is allowed that some $\boldsymbol{x}_0$ are fixed ($\boldsymbol{\Lambda}$=0) and others are stochastic.

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 $\mathbf{R}$ might be time-varying or one may need to include a time-dependent weighting on $\mathbf{R}$[6]. The same applies to $\sum_1^T \frac{1}{2}\log|\mathbf{Q}|$.

All bolded elements are column vectors (lower case) and matrices (upper case). $\mathbf{A}^\top$ is the transpose of matrix $\mathbf{A}$, $\mathbf{A}^{-1}$ is the inverse of $\mathbf{A}$, and $|\mathbf{A}|$ is the determinant of $\mathbf{A}$. Parameters are non-italic while elements that are slanted are realizations of a random variable ($\boldsymbol{x}$ and $\boldsymbol{y}$ are slated)[7]

1.3 Missing values

In Shumway and Stoffer and other presentations of the EM algorithm for MARSS models (Shumway and Stoffer, 2006; Zuur et al., 2003), the missing values case is treated separately from the non-missing values case. In these derivations, a series of modifications are given for the EM update equations when there are missing values. In my derivation, I present the missing values treatment differently, and there is only one set of update equations and these equations apply in both the missing values and non-missing values cases. My derivation does this by keeping $\mathrm{E}[\boldsymbol{Y}_t|\text{data}]$ and $\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\text{data}]$ in the update equations (much like $\mathrm{E}[\boldsymbol{X}_t|\text{data}]$ is kept in the equations) while Shumway and Stoffer replace these expectations involving $\boldsymbol{Y}_t$ by their values, which depend on whether or not the data are a complete observation of $\boldsymbol{Y}_t$ with no missing values. Section 6 shows how to compute the expectations involving $\boldsymbol{Y}_t$ when the data are an incomplete observation of $\boldsymbol{Y}_t$.

[6] 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.

[7] 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 $\boldsymbol{Y}$. 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):

$$\Theta_{j+1} = \arg\max_{\Theta} \int_{\boldsymbol{x}} \int_{\boldsymbol{y}} \log \boldsymbol{L}(\boldsymbol{x}, \boldsymbol{y}; \Theta) f(\boldsymbol{x}, \boldsymbol{y} | \boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j) d\boldsymbol{x} d\boldsymbol{y} \tag{8}$$

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

$$\alpha_{j+1} = \arg\max_{\alpha} \int_{\boldsymbol{x}} \int_{\boldsymbol{y}} \log \boldsymbol{L}(\boldsymbol{x}, \boldsymbol{y}, \beta_j; \alpha) f(\boldsymbol{x}, \boldsymbol{y} | \boldsymbol{Y}(1) = \boldsymbol{y}(1), \alpha_j, \beta_j) d\boldsymbol{x} d\boldsymbol{y} \tag{9}$$

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

Expectation step The integral that appears in equation 8 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 11 on using Monte Carlo initialization to ensure that the global maximum is found.

Dividing the parameters into parts: Above the parameter set is written as Θ. However Θ is composed of multiple components: $\mathbf{B}$, $\mathbf{u}$, $\mathbf{R}$, etc. The EM iteration j is broken into subparts for each parameter matrix and both the maximization and expectation steps are done for each part. For example, the expectation step is run with parameters $\{\mathbf{B}_j, \mathbf{u}_j, \mathbf{R}_j, \dots\}$ and then $\mathbf{B}$ is updated to $\mathbf{B}_{j+1}$ with the maximization step. The expectation step is run with parameters $\{\mathbf{B}_{j+1}, \mathbf{u}_j, \mathbf{R}_j, \dots\}$ and the $\mathbf{u}$is updated to $\mathbf{u}_{j+1}$ with the maximization step. The expectation step is run with parameters $\{\mathbf{B}_{j+1}, \mathbf{u}_{j+1}, \mathbf{R}_j, \dots\}$ and the $\mathbf{R}$ is updated. This is continued until all parameters in Θ are updated and that completes the $j+1$ update.

2.1 The expected log-likelihood function

The function that is maximized in the "M" step is the expected value of the log-likelihood function. This expectation is conditioned on two things: 1) the observed $\boldsymbol{Y}$'s which are denoted $\boldsymbol{Y}(1)$ and which are equal to the fixed values $\boldsymbol{y}(1)$ and 2) the parameter set Θ_j. Note that since there may be missing values in the data, $\boldsymbol{Y}(1)$ can be a subset of $\boldsymbol{Y}$, that is, only some $\boldsymbol{Y}$ have a corresponding $\boldsymbol{y}$ value at time t. Mathematically what we are doing is $\mathrm{E}_{\mathbf{XY}}[g(\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j]$. This is a multivariate conditional expectation because $\boldsymbol{X}, \boldsymbol{Y}$ is multivariate (a $m \times n \times T$ vector). The function $g(\Theta)$ that we are taking the expectation of is $\log \boldsymbol{L}(\boldsymbol{Y}, \boldsymbol{X}; \Theta)$. Note that $g(\Theta)$ is a random variable involving the random variables, $\boldsymbol{X}$ and $\boldsymbol{Y}$, while $\log \boldsymbol{L}(\boldsymbol{y}, \boldsymbol{x}; \Theta)$ 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. The goal is to find the Θ that maximize Ψ and this becomes the new Θ for the $j+1$ iteration of the EM algorithm. The equations to compute the new Θ are termed the update equations. Using the log likelihood equation 6 and expanding out all the terms, we can write out Ψ

in verbose form as:

$$
\begin{aligned}
&\mathrm{E}_{\mathbf{XY}}[\log \boldsymbol{L}(\boldsymbol{Y},\boldsymbol{X};\Theta);\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j] = \Psi = \\
&-\frac{1}{2}\sum_{1}^{T}\Big(\mathrm{E}[\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\boldsymbol{Y}_t] - \mathrm{E}[\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t] - \mathrm{E}[(\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\boldsymbol{Y}_t] - \mathrm{E}[\mathbf{a}^\top\mathbf{R}^{-1}\boldsymbol{Y}_t] - \mathrm{E}[\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\mathbf{a}] \\
&+ \mathrm{E}[(\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t] + \mathrm{E}[\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t] + \mathrm{E}[(\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\mathbf{a}] + \mathrm{E}[\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{a}]\Big) - \frac{T}{2}\log|\mathbf{R}| \\
&-\frac{1}{2}\sum_{1}^{T}\Big(\mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\boldsymbol{X}_t] - \mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}] - \mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\boldsymbol{X}_t] \\
&- \mathrm{E}[\mathbf{u}^\top\mathbf{Q}^{-1}\boldsymbol{X}_t] - \mathrm{E}[\boldsymbol{X}_t^\top\mathbf{Q}^{-1}\mathbf{u}] + \mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}] \\
&+ \mathrm{E}[\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{X}_{t-1}] + \mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{u}] + \mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{u}\Big) - \frac{T}{2}\log|\mathbf{Q}| \\
&-\frac{1}{2}\Big(\mathrm{E}[\boldsymbol{X}_0^\top\boldsymbol{V}_0^{-1}\boldsymbol{X}_0] - \mathrm{E}[\boldsymbol{\xi}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{X}_0] - \mathrm{E}[\boldsymbol{X}_0^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}] + \boldsymbol{\xi}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}\Big) - \frac{1}{2}\log|\boldsymbol{\Lambda}| - \frac{n}{2}\log\pi
\end{aligned}
\tag{10}
$$

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

Keep in mind that Θ and Θ_j are different. Θ is a parameter appearing in function $g(\boldsymbol{X},\boldsymbol{Y},\Theta)$, i.e., the parameters in equation 6. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random variables which means that $g(\boldsymbol{X},\boldsymbol{Y},\Theta)$ is a random variable. We take the expectation of $g(\boldsymbol{X},\boldsymbol{Y},\Theta)$, meaning we take integral over the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$. We need to specify what that distribution is and the conditioning on Θ_j (meaning the Θ_j appearing to the right of the | in $\mathrm{E}[g()|\Theta_j]$) is specifying this distribution. This conditioning affects the value of the expectation of $g(\boldsymbol{X},\boldsymbol{Y},\Theta)$, but it does not affect the value of Θ, which are the $\mathbf{R}$, $\mathbf{Q}$, $\mathbf{u}$, etc. values on the right side of equation 10. We will first take the expectation of $g(\boldsymbol{X},\boldsymbol{Y},\Theta)$ conditioned on Θ_j (using integration) and then take the differential of that expectation with respect to Θ.

2.2 The expectations used in the derivation

The following expectations appear frequently in the update equations and are given special names[8]:

$$\widetilde{\mathbf{x}}_t = \mathrm{E}_{\mathbf{XY}}[\boldsymbol{X}_t|\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j] \tag{11a}$$
$$\widetilde{\mathbf{y}}_t = \mathrm{E}_{\mathbf{XY}}[\boldsymbol{Y}_t|\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j] \tag{11b}$$
$$\widetilde{\mathbf{P}}_t = \mathrm{E}_{\mathbf{XY}}[\boldsymbol{X}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j] \tag{11c}$$
$$\widetilde{\mathbf{P}}_{t,t-1} = \mathrm{E}_{\mathbf{XY}}[\boldsymbol{X}_t\boldsymbol{X}_{t-1}^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j] \tag{11d}$$
$$\widetilde{\boldsymbol{V}}_t = \mathrm{var}_{XY}[\boldsymbol{X}_t|\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j] = \widetilde{\mathbf{P}}_t - \widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^\top \tag{11e}$$
$$\widetilde{\mathbf{O}}_t = \mathrm{E}_{\mathbf{XY}}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j] \tag{11f}$$
$$\widetilde{\boldsymbol{W}}_t = \mathrm{var}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j] = \widetilde{\mathbf{O}}_t - \widetilde{\mathbf{y}}_t\widetilde{\mathbf{y}}_t^\top \tag{11g}$$
$$\widetilde{\mathbf{y}\mathbf{x}}_t = \mathrm{E}_{\mathbf{XY}}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j] \tag{11h}$$
$$\widetilde{\mathbf{y}\mathbf{x}}_{t,t-1} = \mathrm{E}_{\mathbf{XY}}[\boldsymbol{Y}_t\boldsymbol{X}_{t-1}^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1),\Theta_j] \tag{11i}$$

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 11e and 11g arise from the computational formula for variance and covariance:

$$\mathrm{var}[X] = \mathrm{E}[XX^\top] - \mathrm{E}[X]\,\mathrm{E}[X]^\top \tag{12}$$
$$\mathrm{cov}[X,Y] = \mathrm{E}[XY^\top] - \mathrm{E}[X]\,\mathrm{E}[Y]^\top. \tag{13}$$

Section 6 shows how to compute the expectations in equation 11.

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

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

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

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, i.e., the unconstrained case. These are similar to the update equations one will see in Shumway and Stoffer (2006). Section 5 shows the update equations when there are unestimated (fixed) or estimated but shared values in the parameter matrices, i.e., the constrained update equations.

To derive the update equations, one must find the Θ, where Θ is comprised of the MARSS parameters $\mathbf{B}$, $\mathbf{u}$, $\mathbf{Q}$, $\mathbf{Z}$, $\mathbf{a}$, $\mathbf{R}$, $\boldsymbol{\xi}$, and $\boldsymbol{\Lambda}$, that maximizes Ψ (equation 10) by partial differentiation of Ψ with respect to Θ. However, I will be using the EM equation where one maximizes each parameter matrix in Θ one-by-one (equation 9). In this case, the parameters that are not being maximized are fixed (and set at their current iteration value), 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 $u_{i,j}$ in matrix $\mathbf{u}$. The idea is to single out those terms in equation 10 that involve $u_{i,j}$ (say), differentiate by $u_{i,j}$, set this to zero and solve for $u_{i,j}$. This gives the new $u_{i,j}$ that maximizes the partial derivative with respect to $u_{i,j}$ of the expected log-likelihood. Matrix calculus gives us a way to jointly maximize Ψ with respect to all elements (not just element i, j) in a parameter matrix.

Note, see the comments on the EM algorithm implementation (Section 2) when the parameter set Θ is broken into parts (e.g., $\mathbf{B}$, $\mathbf{u}$, $\mathbf{Q}$, etc.). In the implementation of the algorithm, one updates the Θ parts sequentially and the expectation step is re-run with the new Θ at each step (meaning the Kalman smoother is re-run with the updated parameters). Thus the algorithm is applied as follows (order that the parameters are updated is unimportant): E-step with $\{\mathbf{B}_j, \mathbf{u}_j, \mathbf{Q}_j, etc.\}$, M-step updates $\mathbf{B}_j$ to $\mathbf{B}_{j+1}$, E-step with $\{\mathbf{B}_{j+1}, \mathbf{u}_j, \mathbf{Q}_j, etc.\}$, M-step updates $\mathbf{u}_j$ to $\mathbf{u}_{j+1}$, E-step with $\{\mathbf{B}_{j+1}, \mathbf{u}_{j+1}, \mathbf{Q}_j, etc.\}$, M-step updates $\mathbf{Q}_j$ to $\mathbf{Q}_{j+1}$, continuing until all parameters are updates which completes the $j+1$ update.

3.1 Matrix calculus need for the derivation

A number of derivatives of a scalar with respect to vectors and matrices will be needed in the derivation and are shown in table 2. 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

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

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

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

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. The partial derivatives of a column vector $\mathbf{a}$ with respect to a column vector $\mathbf{b}$.

$$\frac{\partial \Psi}{\partial \mathbf{B}} = \begin{bmatrix} \frac{\partial a_1}{\partial b_1} & \frac{\partial a_1}{\partial b_2} & \cdots & \frac{\partial a_1}{\partial b_m} \\ \frac{\partial a_2}{\partial b_1} & \frac{\partial \Psi}{\partial b_2} & \cdots & \frac{\partial a_2}{\partial b_m} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial a_n}{\partial b_1} & \frac{\partial a_n}{\partial b_2} & \cdots & \frac{\partial a_n}{\partial b_m} \end{bmatrix}$$

In table 2, both the vectorized and non-vectorized versions are shown. The vectorized version of a matrix $\mathbf{D}$ with dimension $n \times m$ is

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

3.2 The update equation for u (unconstrained)

Take the partial derivative of Ψ with respect to $\mathbf{u}$, which is a $m \times 1$ matrix. 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 10 with respect to $\mathbf{u}$:

$$\begin{aligned} \partial \Psi / \partial \mathbf{u} = -\frac{1}{2} \sum_{t=1}^{T} \Big(& - \partial(\,\mathrm{E}[\boldsymbol{X}_t^\top \mathbf{Q}^{-1} \mathbf{u}]) / \partial \mathbf{u} - \partial(\,\mathrm{E}[\mathbf{u}^\top \mathbf{Q}^{-1} \boldsymbol{X}_t]) / \partial \mathbf{u} \\ & + \partial(\,\mathrm{E}[(\mathbf{B}\boldsymbol{X}_{t-1})^\top \mathbf{Q}^{-1} \mathbf{u}]) / \partial \mathbf{u} + \partial(\,\mathrm{E}[\mathbf{u}^\top \mathbf{Q}^{-1} \mathbf{B} \boldsymbol{X}_{t-1}]) / \partial \mathbf{u} + \partial(\mathbf{u}^\top \mathbf{Q}^{-1} \mathbf{u}) / \partial \mathbf{u} \Big) \end{aligned} \tag{21}$$

The parameters can be moved out of the expectations and then the matrix derivative relations (table 2) are used to take the derivative.

$$\partial \Psi / \partial \mathbf{u} = -\frac{1}{2} \sum_{t=1}^{T} \Big(- \,\mathrm{E}[\boldsymbol{X}_t]^\top \mathbf{Q}^{-1} - \,\mathrm{E}[\boldsymbol{X}_t]^\top \mathbf{Q}^{-1} + (\mathbf{B}\,\mathrm{E}[\boldsymbol{X}_{t-1}])^\top \mathbf{Q}^{-1} + (\mathbf{B}\,\mathrm{E}[\boldsymbol{X}_{t-1}])^\top \mathbf{Q}^{-1} + 2\mathbf{u}^\top \mathbf{Q}^{-1} \Big) \tag{22}$$

Table 2: Derivatives of a scalar with respect to vectors and matrices. In the following a is a scalar (unrelated to $\mathbf{a}$), $\mathbf{a}$ and $\mathbf{c}$ are $n \times 1$ column vectors, $\mathbf{b}$ and $\mathbf{d}$ are $m \times 1$ column vectors, $\mathbf{D}$ is a $n \times m$ matrix, $\mathbf{C}$ is a $n \times n$ matrix, and $\mathbf{A}$ is a diagonal $n \times n$ matrix (0s on the off-diagonals). $\mathbf{C}^{-1}$ is the inverse of $\mathbf{C}$, $\mathbf{C}^\top$ is the transpose of $\mathbf{C}$, $\mathbf{C}^{-\top} = \left(\mathbf{C}^{-1}\right)^\top = \left(\mathbf{C}^\top\right)^{-1}$, and $|\mathbf{C}|$ is the determinant of $\mathbf{C}$. Note, all the numerators in the differentials on the far left reduce to scalars. Although the matrix names may be the same as those of matrices referred to in the text, the matrices in this table are dummy matrices used to show the matrix derivative relations.

$$\partial(\mathbf{f}^\top\mathbf{g})/\partial\mathbf{a} = \mathbf{f}^\top\partial\mathbf{g}/\partial\mathbf{a} + \mathbf{g}^\top\partial\mathbf{f}/\partial\mathbf{a}$$
$\mathbf{f} = f(\mathbf{a})$ and $\mathbf{g}{=}g(\mathbf{a})$ are $m \times 1$ column vectors and functions of $\mathbf{a}$. (14)
$\partial a/\partial\mathbf{a} = \frac{1}{n}\partial a/\partial\mathbf{g}\,\partial\mathbf{g}/\partial\mathbf{a}$
$\partial\mathbf{f}/\partial\mathbf{a} = \frac{1}{m}\partial\mathbf{f}/\partial\mathbf{g}\,\partial\mathbf{g}/\partial\mathbf{a}$

$$\partial(\mathbf{a}^\top\mathbf{c})/\partial\mathbf{a} = \partial(\mathbf{c}^\top\mathbf{a})/\partial\mathbf{a} = \mathbf{c}^\top$$
$$\partial\mathbf{a}/\partial\mathbf{a} = \partial(\mathbf{a}^\top)/\partial\mathbf{a} = \mathbf{I}_n \quad (15)$$

$$\partial(\mathbf{a}^\top\mathbf{D}\mathbf{b})/\partial\mathbf{D} = \partial(\mathbf{b}^\top\mathbf{D}^\top\mathbf{a})/\partial\mathbf{D} = \mathbf{b}\mathbf{a}^\top$$
$$\partial(\mathbf{a}^\top\mathbf{D}\mathbf{b})/\partial\,\mathrm{vec}(\mathbf{D}) = \partial(\mathbf{b}^\top\mathbf{D}^\top\mathbf{a})/\partial\,\mathrm{vec}(\mathbf{D}) = \big(\mathrm{vec}(\mathbf{b}\mathbf{a}^\top)\big)^\top \quad (16)$$

$\mathbf{C}$ is invertible.
$$\partial(\log|\mathbf{C}|)/\partial\mathbf{C} = -\partial(\log|\mathbf{C}^{-1}|)/\partial\mathbf{C} = (\mathbf{C}^\top)^{-1} = \mathbf{C}^{-\top}$$
$$\partial(\log|\mathbf{C}|)/\partial\,\mathrm{vec}(\mathbf{C}) = \big(\mathrm{vec}(\mathbf{C}^{-\top})\big)^\top \quad (17)$$
If $\mathbf{C}$ is also symmetric and $\mathbf{B}$ is not a function of $\mathbf{C}$.
$$\partial(\log|\mathbf{C}^\top\mathbf{B}\mathbf{C}|)/\partial\mathbf{C} = 2\mathbf{C}^{-1}$$
$$\partial(\log|\mathbf{C}^\top\mathbf{B}\mathbf{C}|)/\partial\,\mathrm{vec}(\mathbf{C}) = 2\big(\mathrm{vec}(\mathbf{C}^{-1})\big)^\top$$

$$\partial(\mathbf{b}^\top\mathbf{D}^\top\mathbf{C}\mathbf{D}\mathbf{d})/\partial\mathbf{D} = \mathbf{d}\mathbf{b}^\top\mathbf{D}^\top\mathbf{C} + \mathbf{b}\mathbf{d}^\top\mathbf{D}^\top\mathbf{C}^\top$$
$$\partial(\mathbf{b}^\top\mathbf{D}^\top\mathbf{C}\mathbf{D}\mathbf{d})/\partial\,\mathrm{vec}(\mathbf{D}) = \big(\mathrm{vec}(\mathbf{d}\mathbf{b}^\top\mathbf{D}^\top\mathbf{C} + \mathbf{b}\mathbf{d}^\top\mathbf{D}^\top\mathbf{C}^\top)\big)^\top \quad (18)$$
If $\mathbf{b} = \mathbf{d}$ and $\mathbf{C}$ is symmetric then the sum reduces to $2\mathbf{b}\mathbf{b}^\top\mathbf{D}^\top\mathbf{C}$

$$\partial(\mathbf{a}^\top\mathbf{C}\mathbf{a})/\partial\mathbf{a} = \partial(\mathbf{a}\mathbf{C}^\top\mathbf{a}^\top)/\partial\mathbf{a} = 2\mathbf{a}^\top\mathbf{C} \quad (19)$$

$$\partial(\mathbf{a}^\top\mathbf{C}^{-1}\mathbf{c})/\partial\mathbf{C} = -\mathbf{C}^{-1}\mathbf{a}\mathbf{c}^\top\mathbf{C}^{-1}$$
$$\partial(\mathbf{a}^\top\mathbf{C}^{-1}\mathbf{c})/\partial\,\mathrm{vec}(\mathbf{C}) = -\big(\mathrm{vec}(\mathbf{C}^{-1}\mathbf{a}\mathbf{c}^\top\mathbf{C}^{-1})\big)^\top \quad (20)$$

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

$$\partial\Psi/\partial\mathbf{u} = \sum_{t=1}^{T} \left(\mathrm{E}[\boldsymbol{X}_t]^\top \mathbf{Q}^{-1} - \mathrm{E}[\boldsymbol{X}_{t-1}]^\top \mathbf{B}^\top \mathbf{Q}^{-1} - \mathbf{u}^\top \mathbf{Q}^{-1} \right) \tag{23}$$

Set the left side to zero (a $p \times m$ matrix of zeros) and transpose the whole equation. $\mathbf{Q}^{-1}$ cancels out[9] by multiplying on the left by $\mathbf{Q}$ (left since the whole equation was just transposed), giving

$$\mathbf{0} = \sum_{t=1}^{T} \left(\mathrm{E}[\boldsymbol{X}_t] - \mathbf{B}\,\mathrm{E}[\boldsymbol{X}_{t-1}] - \mathbf{u} \right) = \sum_{t=1}^{T} \left(\mathrm{E}[\boldsymbol{X}_t] - \mathbf{B}\,\mathrm{E}[\boldsymbol{X}_{t-1}] \right) - \mathbf{u} \tag{24}$$

Solving for $\mathbf{u}$ and replacing the expectations with their names from equation 11, gives us the new $\mathbf{u}$ that maximizes Ψ,

$$\mathbf{u}_{j+1} = \frac{1}{T} \sum_{t=1}^{T} \left(\widetilde{\mathbf{x}}_t - \mathbf{B}\widetilde{\mathbf{x}}_{t-1} \right) \tag{25}$$

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

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

After pulling the constants out of the expectations, we use relations 16 and 18 to take the derivative and note that $\mathbf{Q}^{-1} = (\mathbf{Q}^{-1})^\top$:

$$\begin{aligned}
\partial\Psi/\partial\mathbf{B} = -\frac{1}{2} \sum_{t=1}^{T} \Bigg(& - \mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_t^\top]\mathbf{Q}^{-1} - \mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_t^\top]\mathbf{Q}^{-1} \\
& + 2\,\mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top]\mathbf{B}^\top \mathbf{Q}^{-1} + \mathrm{E}[\boldsymbol{X}_{t-1}]\mathbf{u}^\top \mathbf{Q}^{-1} + \mathrm{E}[\boldsymbol{X}_{t-1}]\mathbf{u}^\top \mathbf{Q}^{-1} \Bigg)
\end{aligned} \tag{27}$$

This can be reduced to

$$\partial\Psi/\partial\mathbf{B} = -\frac{1}{2} \sum_{t=1}^{T} \left(-2\,\mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_t^\top]\mathbf{Q}^{-1} + 2\,\mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top]\mathbf{B}^\top \mathbf{Q}^{-1} + 2\,\mathrm{E}[\boldsymbol{X}_{t-1}]\mathbf{u}^\top \mathbf{Q}^{-1} \right) \tag{28}$$

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

$$\begin{aligned}
\mathbf{0} &= \sum_{t=1}^{T} \left(\mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_{t-1}^\top] - \mathbf{B}\,\mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top] - \mathbf{u}\,\mathrm{E}[\boldsymbol{X}_{t-1}^\top] \right) \\
&= \sum_{t=1}^{T} \left(\widetilde{\mathbf{P}}_{t,t-1} - \mathbf{B}\widetilde{\mathbf{P}}_{t-1} - \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top \right)
\end{aligned} \tag{29}$$

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

The last line replaced the expectations with their names shown in equation 11. Solving for $\mathbf{B}$ and noting that $\widetilde{\mathbf{P}}_{t-1}$ is like a variance-covariance matrix and is invertible, gives us the new $\mathbf{B}$ that maximizes Ψ,

$$\mathbf{B}_{j+1} = \left(\sum_{t=1}^{T} \left(\widetilde{\mathbf{P}}_{t,t-1} - \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^{\top} \right) \right) \left(\sum_{t=1}^{T} \widetilde{\mathbf{P}}_{t-1} \right)^{-1} \tag{30}$$

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 $\mathbf{B}$,

$$\mathbf{B}_{i,j+1} = \left(\sum_{t=1}^{T} \left(\widetilde{\mathbf{P}}_{t,t-1} - \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^{\top} \right) \right)_i \left(\sum_{t=1}^{T} \widetilde{\mathbf{P}}_{t-1} \right)_i^{-1} \tag{31}$$

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 31.

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 (equation 10). 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 $\partial(\,\mathrm{E}[h(\boldsymbol{X}_t, \mathbf{Q})])/\partial\mathbf{Q} = \mathrm{E}[\partial(h(\boldsymbol{X}_t, \mathbf{Q}))/\partial\mathbf{Q}]$.

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

The relations (20) and (17) 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 20 and it cancels out the minus in front of the initial $-1/2$.

$$\begin{aligned} \partial\Psi/\partial\mathbf{Q} = \frac{1}{2}\sum_{t=1}^{T} \mathbf{Q}^{-1} \Bigg(& \mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_t^{\top}] - \mathrm{E}[\boldsymbol{X}_t(\mathbf{B}\boldsymbol{X}_{t-1})^{\top}] - \mathrm{E}[\mathbf{B}\boldsymbol{X}_{t-1}\boldsymbol{X}_t^{\top}] - \mathrm{E}[\boldsymbol{X}_t\mathbf{u}^{\top}] - \mathrm{E}[\mathbf{u}\boldsymbol{X}_t^{\top}] \\ & + \mathrm{E}[\mathbf{B}\boldsymbol{X}_{t-1}(\mathbf{B}\boldsymbol{X}_{t-1})^{\top}] + \mathrm{E}[\mathbf{B}\boldsymbol{X}_{t-1}\mathbf{u}^{\top}] + \mathrm{E}[\mathbf{u}(\mathbf{B}\boldsymbol{X}_{t-1})^{\top}] + \mathbf{u}\mathbf{u}^{\top} \Bigg) \mathbf{Q}^{-1} - \frac{T}{2}\mathbf{Q}^{-1} \end{aligned} \tag{33}$$

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

$$\begin{aligned}\partial\Psi/\partial\mathbf{Q} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{Q}^{-1}\Big(&\operatorname{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top] - \operatorname{E}[\boldsymbol{X}_t\boldsymbol{X}_{t-1}^\top]\mathbf{B}^\top - \mathbf{B}\operatorname{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_t^\top] - \operatorname{E}[\boldsymbol{X}_t]\mathbf{u}^\top - \mathbf{u}\operatorname{E}[\boldsymbol{X}_t^\top] \\ &+ \mathbf{B}\operatorname{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top]\mathbf{B}^\top + \mathbf{B}\operatorname{E}[\boldsymbol{X}_{t-1}]\mathbf{u}^\top + \mathbf{u}\operatorname{E}[\boldsymbol{X}_{t-1}^\top]\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\Big)\mathbf{Q}^{-1} - \frac{T}{2}\mathbf{Q}^{-1}\end{aligned} \tag{34}$$

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

$$\begin{aligned}\partial\Psi/\partial\mathbf{Q} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{Q}^{-1}\Big(&\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}\widetilde{\mathbf{x}}_t^\top \\ &+ \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\Big)\mathbf{Q}^{-1} - \frac{T}{2}\mathbf{Q}^{-1}\end{aligned} \tag{35}$$

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:

$$T\mathbf{Q} = \sum_{t=1}^{T}\left(\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}\widetilde{\mathbf{x}}_t^\top + \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\right) \tag{36}$$

This gives us the new $\mathbf{Q}$ that maximizes Ψ,

$$\begin{aligned}\mathbf{Q}_{j+1} = \frac{1}{T}\sum_{t=1}^{T}\Big(&\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}\widetilde{\mathbf{x}}_t^\top \\ &+ \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\Big)\end{aligned} \tag{37}$$

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

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

In this case,

$$\begin{aligned}\mathbf{Q}_{i,j+1} = \frac{1}{T}\sum_{t=1}^{T}\Big(&\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}\widetilde{\mathbf{x}}_t^\top \\ &+ \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}\widetilde{\mathbf{x}}_{t-1}^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\Big)_i\end{aligned} \tag{38}$$

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 38.

By the way, $\mathbf{Q}$ is never really unconstrained since it is a variance-covariance matrix and the upper and lower triangles are shared. However, because the shared values are only the symmetric values in the matrix, the derivation still works even though it's technically incorrect (Henderson and Searle, 1979). The constrained update equation for $\mathbf{Q}$ shown in section 5.8 explicitly deals with the shared lower and upper triangles.

3.5 Update equation for a (unconstrained)

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

$$\begin{aligned}\partial\Psi/\partial\mathbf{a} = -\frac{1}{2}\sum_{t=1}^{T}\Big(&- \partial(\,\mathrm{E}[\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\mathbf{a}])/\partial\mathbf{a} - \partial(\,\mathrm{E}[\mathbf{a}^\top\mathbf{R}^{-1}\boldsymbol{Y}_t])/\partial\mathbf{a} \\ &+ \partial(\,\mathrm{E}[(\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\mathbf{a}])/\partial\mathbf{a} + \partial(\,\mathrm{E}[\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t])/\partial\mathbf{a} + \partial(\,\mathrm{E}[\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{a}])/\partial\mathbf{a}\Big)\end{aligned} \tag{39}$$

The expectations around constants can be dropped[10]. Using relations (15) and (19) and using $\mathbf{R}^{-1} = (\mathbf{R}^{-1})^\top$, we have then

$$\partial\Psi/\partial\mathbf{a} = -\frac{1}{2}\sum_{t=1}^{T}\Big(- \,\mathrm{E}[\boldsymbol{Y}_t^\top\mathbf{R}^{-1}] - \,\mathrm{E}[\boldsymbol{Y}_t^\top\mathbf{R}^{-1}] + \,\mathrm{E}[(\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}] + \,\mathrm{E}[(\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}] + 2\mathbf{a}^\top\mathbf{R}^{-1}\Big) \tag{40}$$

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

$$\partial\Psi/\partial\mathbf{a} = \sum_{t=1}^{T}\Big(\,\mathrm{E}[\boldsymbol{Y}_t]^\top\mathbf{R}^{-1} - \,\mathrm{E}[\boldsymbol{X}_t]^\top\mathbf{Z}^\top\mathbf{R}^{-1} - \mathbf{a}^\top\mathbf{R}^{-1}\Big) \tag{41}$$

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

$$\mathbf{0} = \sum_{t=1}^{T}\big(\,\mathrm{E}[\boldsymbol{Y}_t] - \mathbf{Z}\,\mathrm{E}[\boldsymbol{X}_t] - \mathbf{a}\big) = \sum_{t=1}^{T}\big(\widetilde{\mathbf{y}}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a}\big) \tag{42}$$

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

$$\mathbf{a}_{j+1} = \frac{1}{T}\sum_{t=1}^{T}(\widetilde{\mathbf{y}}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t) \tag{43}$$

3.6 The update equation for Z (unconstrained)

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

$$\begin{aligned}&\partial\Psi/\partial\mathbf{Z} = \text{(note } \partial\mathbf{Z} \text{ is } m \times n \text{ while } \mathbf{Z} \text{ is } n \times m\text{)} \\ &-\frac{1}{2}\sum_{t=1}^{T}\Big(- \,\mathrm{E}[\partial(\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{Z}] - \,\mathrm{E}[\partial((\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\boldsymbol{Y}_t)/\partial\mathbf{Z}] + \,\mathrm{E}[\partial((\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{Z}] \\ &+ \,\mathrm{E}[\partial((\mathbf{Z}\boldsymbol{X}_t)^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{Z}] + \,\mathrm{E}[\partial(\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{Z}]\Big) \\ &= -\frac{1}{2}\sum_{t=1}^{T}\Big(- \,\mathrm{E}[\partial(\boldsymbol{Y}_t^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{Z}] - \,\mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Z}^\top\mathbf{R}^{-1}\boldsymbol{Y}_t)/\partial\mathbf{Z}] + \,\mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{Z}] \\ &+ \,\mathrm{E}[\partial(\boldsymbol{X}_t^\top\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{Z}] + \,\mathrm{E}[\partial(\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{X}_t)/\partial\mathbf{Z}]\Big)\end{aligned} \tag{44}$$

Using the matrix derivative relations (table 2) and using $\mathbf{R}^{-1} = (\mathbf{R}^{-1})^\top$, we get

$$\begin{aligned}\partial\Psi/\partial\mathbf{Z} = -\frac{1}{2}\sum_{t=1}^{T}\Big(&- \,\mathrm{E}[\boldsymbol{X}_t\boldsymbol{Y}_t^\top\mathbf{R}^{-1}] - \,\mathrm{E}[\boldsymbol{X}_t\boldsymbol{Y}_t^\top\mathbf{R}^{-1}] \\ &+ 2\,\mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top\mathbf{Z}^\top\mathbf{R}^{-1}] + \,\mathrm{E}[\boldsymbol{X}_{t-1}\mathbf{a}^\top\mathbf{R}^{-1}] + \,\mathrm{E}[\boldsymbol{X}_t\mathbf{a}^\top\mathbf{R}^{-1}]\Big)\end{aligned} \tag{45}$$

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

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

$$\partial\Psi/\partial\mathbf{Z} = \sum_{t=1}^{T}\Big(\mathrm{E}[\boldsymbol{X}_t\boldsymbol{Y}_t^\top]\mathbf{R}^{-1} - \mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top]\mathbf{Z}^\top\mathbf{R}^{-1} - \mathrm{E}[\boldsymbol{X}_t]\mathbf{a}^\top\mathbf{R}^{-1}\Big) \tag{46}$$

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

$$\mathbf{0} = \sum_{t=1}^{T}\big(\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top] - \mathbf{Z}\,\mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top] - \mathbf{a}\,\mathrm{E}[\boldsymbol{X}_t^\top]\big) = \sum_{t=1}^{T}\big(\widetilde{\mathbf{y}\mathbf{x}}_t - \mathbf{Z}\widetilde{\mathbf{P}}_t - \mathbf{a}\widetilde{\mathbf{x}}_t^\top\big) \tag{47}$$

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

$$\mathbf{Z}_{j+1} = \Big(\sum_{t=1}^{T}\big(\widetilde{\mathbf{y}\mathbf{x}}_t - \mathbf{a}\widetilde{\mathbf{x}}_t^\top\big)\Big)\Big(\sum_{t=1}^{T}\widetilde{\mathbf{P}}_t\Big)^{-1} \tag{48}$$

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.

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

We use relations (20) and (17) 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 20 and cancels out the minus in front of $-1/2$.

$$\begin{aligned}\partial\Psi/\partial\mathbf{R} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{R}^{-1}\Big(&\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top] - \mathrm{E}[\boldsymbol{Y}_t(\mathbf{Z}\boldsymbol{X}_t)^\top] - \mathrm{E}[\mathbf{Z}\boldsymbol{X}_t\boldsymbol{Y}_t^\top] - \mathrm{E}[\boldsymbol{Y}_t\mathbf{a}^\top] - \mathrm{E}[\mathbf{a}\boldsymbol{Y}_t^\top]\\ &+ \mathrm{E}[\mathbf{Z}\boldsymbol{X}_t(\mathbf{Z}\boldsymbol{X}_t)^\top] + \mathrm{E}[\mathbf{Z}\boldsymbol{X}_t\mathbf{a}^\top] + \mathrm{E}[\mathbf{a}(\mathbf{Z}\boldsymbol{X}_t)^\top] + \mathbf{a}\mathbf{a}^\top\Big)\mathbf{R}^{-1} - \frac{T}{2}\mathbf{R}^{-1}\end{aligned} \tag{50}$$

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

$$\begin{aligned}\partial\Psi/\partial\mathbf{R} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{R}^{-1}\Big(&\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top] - \mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top]\mathbf{Z}^\top - \mathbf{Z}\,\mathrm{E}[\boldsymbol{X}_t\boldsymbol{Y}_t^\top] - \mathrm{E}[\boldsymbol{Y}_t]\mathbf{a}^\top - \mathbf{a}\,\mathrm{E}[\boldsymbol{Y}_t^\top]\\ &+ \mathbf{Z}\,\mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top]\mathbf{Z}^\top + \mathbf{Z}\,\mathrm{E}[\boldsymbol{X}_t]\mathbf{a}^\top + \mathbf{a}\,\mathrm{E}[\boldsymbol{X}_t^\top]\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\Big)\mathbf{R}^{-1} - \frac{T}{2}\mathbf{R}^{-1}\end{aligned} \tag{51}$$

We rewrite the partial derivative in terms of expectations:

$$\begin{aligned}\partial\Psi/\partial\mathbf{R} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{R}^{-1}\Big(&\widetilde{\mathbf{O}}_t - \widetilde{\mathbf{y}\mathbf{x}}_t\mathbf{Z}^\top - \mathbf{Z}\widetilde{\mathbf{y}\mathbf{x}}_t^\top - \widetilde{\mathbf{y}}_t\mathbf{a}^\top - \mathbf{a}\widetilde{\mathbf{y}}_t^\top\\ &+ \mathbf{Z}\widetilde{\mathbf{P}}_t\mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{x}}_t\mathbf{a}^\top + \mathbf{a}\widetilde{\mathbf{x}}_t^\top\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\Big)\mathbf{R}^{-1} - \frac{T}{2}\mathbf{R}^{-1}\end{aligned} \tag{52}$$

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 $1/2$.

$$T\mathbf{R} = \sum_{t=1}^{T}\Big(\widetilde{\mathbf{O}}_t - \widetilde{\mathbf{y}\mathbf{x}}_t\mathbf{Z}^\top - \mathbf{Z}\widetilde{\mathbf{y}\mathbf{x}}_t^\top - \widetilde{\mathbf{y}}_t\mathbf{a}^\top - \mathbf{a}\widetilde{\mathbf{y}}_t^\top + \mathbf{Z}\widetilde{\mathbf{P}}_t\mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{x}}_t\mathbf{a}^\top + \mathbf{a}\widetilde{\mathbf{x}}_t^\top\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\Big) \tag{53}$$

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

$$\mathbf{R}_{j+1} = \frac{1}{T}\sum_{t=1}^{T}\left(\widetilde{\mathbf{O}}_t - \widetilde{\mathbf{y}\mathbf{x}}_t\mathbf{Z}^\top - \mathbf{Z}\widetilde{\mathbf{y}\mathbf{x}}_t^\top - \widetilde{\mathbf{y}}_t\mathbf{a}^\top - \mathbf{a}\widetilde{\mathbf{y}}_t^\top + \mathbf{Z}\widetilde{\mathbf{P}}_t\mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{x}}_t\mathbf{a}^\top + \mathbf{a}\widetilde{\mathbf{x}}_t^\top\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\right) \tag{54}$$

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,

$$\mathbf{R}_{i,j+1} = \frac{1}{T}\sum_{t=1}^{T}\left(\widetilde{\mathbf{O}}_t - \widetilde{\mathbf{y}\mathbf{x}}_t\mathbf{Z}^\top - \mathbf{Z}\widetilde{\mathbf{y}\mathbf{x}}_t^\top - \widetilde{\mathbf{y}}_t\mathbf{a}^\top - \mathbf{a}\widetilde{\mathbf{y}}_t^\top + \mathbf{Z}\widetilde{\mathbf{P}}_t\mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{x}}_t\mathbf{a}^\top + \mathbf{a}\widetilde{\mathbf{x}}_t^\top\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\right)_i \tag{55}$$

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 55.

3.8 Update equation for $\boldsymbol{\xi}$ and $\boldsymbol{\Lambda}$ (unconstrained), stochastic initial state

Shumway and Stoffer (2006) and Ghahramani and Hinton (1996) imply in their discussion of the EM algorithm that both $\boldsymbol{\xi}$ and $\boldsymbol{\Lambda}$ can be estimated (though not simultaneously). Harvey (1989), however, discusses that there are only two allowable cases: $\boldsymbol{x}_0$ is treated as fixed ($\boldsymbol{\Lambda} = 0$) and equal to the unknown parameter $\boldsymbol{\xi}$ or $\boldsymbol{x}_0$ is treated as stochastic with a known mean $\boldsymbol{\xi}$ and variance $\boldsymbol{\Lambda}$. For completeness, we show here the update equation in the case of $\boldsymbol{x}_0$ stochastic with unknown mean $\boldsymbol{\xi}$ and variance $\boldsymbol{\Lambda}$ (a case that Harvey (1989) says is not consistent).

We 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.

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

Using relations (15) and (19) and using $\boldsymbol{\Lambda}^{-1} = (\boldsymbol{\Lambda}^{-1})^\top$, we have

$$\partial\Psi/\partial\boldsymbol{\xi} = -\frac{1}{2}\big(-\,\mathrm{E}[\boldsymbol{X}_0^\top\boldsymbol{\Lambda}^{-1}] - \,\mathrm{E}[\boldsymbol{X}_0^\top\boldsymbol{\Lambda}^{-1}] + 2\boldsymbol{\xi}^\top\boldsymbol{\Lambda}^{-1}\big) \tag{57}$$

Pulling the parameters out of the expectations, we get

$$\partial\Psi/\partial\boldsymbol{\xi} = -\frac{1}{2}\big(-2\,\mathrm{E}[\boldsymbol{X}_0^\top]\boldsymbol{\Lambda}^{-1} + 2\boldsymbol{\xi}^\top\boldsymbol{\Lambda}^{-1}\big) \tag{58}$$

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

$$\mathbf{0} = \big(\boldsymbol{\Lambda}^{-1}\,\mathrm{E}[\boldsymbol{X}_0] + \boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}\big) = (\widetilde{\mathbf{x}}_0 - \boldsymbol{\xi}) \tag{59}$$

Thus,

$$\boldsymbol{\xi}_{j+1} = \widetilde{\mathbf{x}}_0 \tag{60}$$

$\widetilde{\mathbf{x}}_0$ 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 $\mathrm{E}[\boldsymbol{X}_0|\boldsymbol{Y}_0 = \boldsymbol{y}_0] \equiv \boldsymbol{\xi}_j$ and $\mathrm{var}(\boldsymbol{X}_0\boldsymbol{X}_0^\top|\boldsymbol{Y}_0 = \boldsymbol{y}_0) \equiv \boldsymbol{\Lambda}_j$ (meaning the filter recursions start with $t = 1$ with $\tilde{x}_{t-1}^{t-1} = \tilde{x}_0^0 = \boldsymbol{\xi}_j$). A similar set of steps gets us to the update equation for $\boldsymbol{\Lambda}$,

$$\boldsymbol{\Lambda}_{j+1} = \widetilde{\boldsymbol{V}}_0 \tag{61}$$

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

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

$$\boldsymbol{\xi}_{j+1} = \widetilde{\mathbf{x}}_1 \tag{62}$$

$$\boldsymbol{\Lambda}_{j+1} = \widetilde{\boldsymbol{V}}_1 \tag{63}$$

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

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

For the case where $\boldsymbol{x}_0$ is treated as fixed, i.e., as another parameter, then there is no $\boldsymbol{\Lambda}$, and we need to maximize $\partial\Psi/\partial\boldsymbol{\xi}$ using the slightly different Ψ shown in equation 7. Now $\boldsymbol{\xi}$ appears in the state equation part of the likelihood.

$$\begin{aligned}
\partial\Psi/\partial\boldsymbol{\xi} &= -\frac{1}{2}\Big(- \mathrm{E}[\partial(\boldsymbol{X}_1^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{\xi})/\partial\boldsymbol{\xi}] - \mathrm{E}[\partial((\mathbf{B}\boldsymbol{\xi})^\top\mathbf{Q}^{-1}\boldsymbol{X}_1)/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{\xi})^\top\mathbf{Q}^{-1}(\mathbf{B}\boldsymbol{\xi}))/\partial\boldsymbol{\xi}] \\
&\quad + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{\xi})^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{\xi})/\partial\boldsymbol{\xi}]\Big) \\
&= -\frac{1}{2}\Big(- \mathrm{E}[\partial(\boldsymbol{X}_1^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{\xi})/\partial\boldsymbol{\xi}] - \mathrm{E}[\partial(\boldsymbol{\xi}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\boldsymbol{X}_1)/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial(\boldsymbol{\xi}^\top\mathbf{B}^\top\mathbf{Q}^{-1}(\mathbf{B}\boldsymbol{\xi}))/\partial\boldsymbol{\xi}] \\
&\quad + \mathrm{E}[\partial(\boldsymbol{\xi}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{\xi})/\partial\boldsymbol{\xi}]\Big)
\end{aligned} \tag{64}$$

After pulling the constants out of the expectations, we use relations (16) and (18) to take the derivative:

$$\partial\Psi/\partial\boldsymbol{\xi} = -\frac{1}{2}\Big(- \mathrm{E}[\boldsymbol{X}_1]^\top\mathbf{Q}^{-1}\mathbf{B} - \mathrm{E}[\boldsymbol{X}_1]^\top\mathbf{Q}^{-1}\mathbf{B} + 2\boldsymbol{\xi}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B} + \mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B} + \mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\Big) \tag{65}$$

This can be reduced to

$$\partial\Psi/\partial\boldsymbol{\xi} = \mathrm{E}[\boldsymbol{X}_1]^\top\mathbf{Q}^{-1}\mathbf{B} - \boldsymbol{\xi}^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B} - \mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B} \tag{66}$$

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.

$$\boldsymbol{\xi} = (\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{Q}^{-1}(\,\mathrm{E}[\boldsymbol{X}_1] - \mathbf{u}) \tag{67}$$

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

$$\boldsymbol{\xi}_{j+1} = (\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{Q}^{-1}(\widetilde{\mathbf{x}}_1 - \mathbf{u}) \tag{68}$$

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

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 7). 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 6) or fixed (equation 7).

If your MARSS model is stationary[11] and your data appear stationary, however, equation 67 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

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

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 $\boldsymbol{\Lambda}$. 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 $\boldsymbol{\Lambda}$ or try fixing $\boldsymbol{\xi}$ and estimating $\boldsymbol{\Lambda}$ (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). A second approach is to estimate $\boldsymbol{x}_1$ as the initial state instead of $\boldsymbol{x}_0$.

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

In some cases, the estimate of $\boldsymbol{x}_0$ from $\boldsymbol{x}_1$ using equation 68 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 can 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 parameter $\boldsymbol{\xi}$. The likelihood is then:

$$\begin{aligned}\log \boldsymbol{L}(\boldsymbol{y}, \boldsymbol{x}; \Theta) = &-\sum_1^T \frac{1}{2}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a})^\top \mathbf{R}^{-1}(\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a}) - \sum_1^T \frac{1}{2}\log|\mathbf{R}| \\ &- \sum_2^T \frac{1}{2}(\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u})^\top \mathbf{Q}^{-1}(\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u}) - \sum_1^T \frac{1}{2}\log|\mathbf{Q}|\end{aligned} \tag{69}$$

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\xi} = &-\frac{1}{2}\Big(- \mathrm{E}[\partial(\boldsymbol{Y}_1^\top \mathbf{R}^{-1}\mathbf{Z}\boldsymbol{\xi})/\partial\boldsymbol{\xi}] - \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{\xi})^\top \mathbf{R}^{-1}\boldsymbol{Y}_1)/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{\xi})^\top \mathbf{R}^{-1}(\mathbf{Z}\boldsymbol{\xi}))/\partial\boldsymbol{\xi}] \\ &+ \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{\xi})^\top \mathbf{R}^{-1}\mathbf{a})/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial(\mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z}\boldsymbol{\xi})/\partial\boldsymbol{\xi}]\Big) \\ &-\frac{1}{2}\Big(- \mathrm{E}[\partial(\boldsymbol{X}_2^\top \mathbf{Q}^{-1}\mathbf{B}\boldsymbol{\xi})/\partial\boldsymbol{\xi}] - \mathrm{E}[\partial((\mathbf{B}\boldsymbol{\xi})^\top \mathbf{Q}^{-1}\boldsymbol{X}_2)/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{\xi})^\top \mathbf{Q}^{-1}(\mathbf{B}\boldsymbol{\xi}))/\partial\boldsymbol{\xi}] \\ &+ \mathrm{E}[\partial((\mathbf{B}\boldsymbol{\xi})^\top \mathbf{Q}^{-1}\mathbf{u})/\partial\boldsymbol{\xi}] + \mathrm{E}[\partial(\mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\boldsymbol{\xi})/\partial\boldsymbol{\xi}]\Big)\end{aligned} \tag{70}$$

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

After pulling the constants out of the expectations, we use relations (16) and (18) to take the derivative:

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\xi} = &-\frac{1}{2}\Big(- \mathrm{E}[\boldsymbol{Y}_1]^\top \mathbf{R}^{-1}\mathbf{Z} - \mathrm{E}[\boldsymbol{Y}_1]^\top \mathbf{R}^{-1}\mathbf{Z} + 2\boldsymbol{\xi}^\top \mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z}\Big) \\ &-\frac{1}{2}\Big(- \mathrm{E}[\boldsymbol{X}_2]^\top \mathbf{Q}^{-1}\mathbf{B} - \mathrm{E}[\boldsymbol{X}_2]^\top \mathbf{Q}^{-1}\mathbf{B} + 2\boldsymbol{\xi}^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B} + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B} + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\Big)\end{aligned} \tag{71}$$

This can be reduced to

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\xi} &= \mathrm{E}[\boldsymbol{Y}_1]^\top \mathbf{R}^{-1}\mathbf{Z} - \boldsymbol{\xi}^\top \mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} - \mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathrm{E}[\boldsymbol{X}_2]^\top \mathbf{Q}^{-1}\mathbf{B} - \boldsymbol{\xi}^\top \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B} - \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B} \\ &= -\boldsymbol{\xi}^\top(\mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B}) + \mathrm{E}[\boldsymbol{Y}_1]^\top \mathbf{R}^{-1}\mathbf{Z} - \mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathrm{E}[\boldsymbol{X}_2]^\top \mathbf{Q}^{-1}\mathbf{B} - \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\end{aligned} \tag{72}$$

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 $\boldsymbol{\xi}$.

$$\boldsymbol{\xi} = (\mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B})^{-1}(\mathbf{Z}^\top \mathbf{R}^{-1}(\,\mathrm{E}[\boldsymbol{Y}_1] - \mathbf{a}) + \mathbf{B}^\top \mathbf{Q}^{-1}(\,\mathrm{E}[\boldsymbol{X}_2] - \mathbf{u})) \tag{73}$$

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

$$\boldsymbol{\xi}_{j+1} = (\mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{B}^\top \mathbf{Q}^{-1}\mathbf{B})^{-1}(\mathbf{Z}^\top \mathbf{R}^{-1}(\widetilde{\mathbf{y}}_1 - \mathbf{a}) + \mathbf{B}^\top \mathbf{Q}^{-1}(\widetilde{\mathbf{x}}_2 - \mathbf{u})) \tag{74}$$

4 The time-varying MARSS model with linear constraints

The first part of this report dealt with the case of a MARSS model (equation 1) where the parameters are time-constant and where all the elements in a parameter matrix are estimated with no constraints. I will now describe the derivation of an EM algorithm to solve a much more general MARSS model (equation 75), which is a time-varying MARSS model where the MARSS parameter matrices are written as a linear equation $\mathbf{f} + \mathbf{Dm}$. This is a very general form of a MARSS model, of which many (most) multivariate autoregressive Gaussian models are a special case. This general MARSS model includes as special cases, MARSS models with covariates (many VARSS models with exogeneous variables), multivariate AR lag-p models and multivariate moving average models, and MARSS models with linear constraints placed on the elements within the model parameters. The objective is to derive one EM algorithm for the whole class, thus a uniform approach to fitting these models.

The time-varying MARSS model is written:

$$\boldsymbol{x}_t = \mathbf{B}_t\boldsymbol{x}_{t-1} + \mathbf{u}_t + \mathbf{G}_t\boldsymbol{w}_t, \text{ where } \boldsymbol{W}_t \sim \text{MVN}(0, \mathbf{Q}_t) \tag{75a}$$

$$\boldsymbol{y}_t = \mathbf{Z}_t\boldsymbol{x}_t + \mathbf{a}_t + \mathbf{H}_t\boldsymbol{v}_t, \text{ where } \boldsymbol{V}_t \sim \text{MVN}(0, \mathbf{R}_t) \tag{75b}$$

$$\boldsymbol{x}_{t_0} = \boldsymbol{\xi} + \mathbf{F}\boldsymbol{l}, \text{ where } t_0 = 0 \text{ or } t_0 = 1 \tag{75c}$$

$$\boldsymbol{L} \sim \text{MVN}(0, \boldsymbol{\Lambda}) \tag{75d}$$

$$\begin{bmatrix}\boldsymbol{w}_t \\ \boldsymbol{v}_t\end{bmatrix} \sim \text{MVN}(0, \Sigma), \quad \Sigma = \begin{bmatrix}\mathbf{Q}_t & 0 \\ 0 & \mathbf{R}_t\end{bmatrix} \tag{75e}$$

This looks quite similar to the previous non-time varying MARSS model, but now the model parameters, $\mathbf{B}$, $\mathbf{u}$, $\mathbf{Q}$, $\mathbf{Z}$, $\mathbf{a}$ and $\mathbf{R}$, have a t subscript and we have a multiplier matrix on the error terms $\boldsymbol{v}_t$, $\boldsymbol{w}_t$, $\boldsymbol{l}$. The $\mathbf{G}_t$ multiplier is $m \times s$, so we now have s state errors instead of m. The $\mathbf{H}_t$ multiplier is $n \times k$, so we now have k observation errors instead of n. The $\mathbf{F}$ multiplier is $m \times j$, so now we can have some initial states (j of them) be stochastic and others be fixed. I assume that appropriate constraints are put on $\mathbf{G}$ and $\mathbf{H}$ so that the resulting MARSS model is not under- or over-constrained[12]. The notation/presentation here was influenced by SJ Koopman's work, esp. Koopman and Ooms (2011) and Koopman (1993), but in these works, $\mathbf{Q}_t$ and $\mathbf{R}_t$ equal $\mathbf{I}$ and the variance-covariance structures are instead specified only by $\mathbf{H}_t$ and $\mathbf{G}_t$. I keep $\mathbf{Q}_t$ and $\mathbf{R}_t$ in my formulation as it seems more intuitive (to me) in the context of the EM algorithm and the required joint-likelihood function.

We can rewrite this MARSS model using vec relationships (table 3):

$$\begin{aligned}\boldsymbol{x}_t &= (\boldsymbol{x}_{t-1}^\top \otimes \mathbf{I}_m)\,\text{vec}(\mathbf{B}_t) + \text{vec}(\mathbf{u}_t) + \mathbf{G}_t\boldsymbol{w}_t, \boldsymbol{W}_t \sim \text{MVN}(0, \mathbf{Q}_t) \\ \boldsymbol{y}_t &= (\boldsymbol{x}_t^\top \otimes \mathbf{I}_n)\,\text{vec}(\mathbf{Z}_t) + \text{vec}(\mathbf{a}_t) + \mathbf{H}_t\boldsymbol{v}_t, \boldsymbol{V}_t \sim \text{MVN}(0, \mathbf{R}_t) \\ \boldsymbol{x}_{t_0} &= \boldsymbol{\xi} + \mathbf{F}\boldsymbol{l}, \boldsymbol{L} \sim \text{MVN}(0, \boldsymbol{\Lambda})\end{aligned} \tag{76}$$

Each model parameter, $\mathbf{B}_t$, $\mathbf{u}_t$, $\mathbf{Q}_t$, $\mathbf{Z}_t$, $\mathbf{a}_t$, and $\mathbf{R}_t$, is written as a time-varying linear model, $\mathbf{f}_t + \mathbf{D}_t\mathbf{m}$, where $\mathbf{f}$ and $\mathbf{D}$ are fully-known (not estimated and no missing values) and $\mathbf{m}$ is a column vector of the estimates elements of the parameter matrix:

$$\begin{aligned}\text{vec}(\mathbf{B}_t) &= \mathbf{f}_{t,b} + \mathbf{D}_{t,b}\boldsymbol{\beta} \\ \text{vec}(\mathbf{u}_t) &= \mathbf{f}_{t,u} + \mathbf{D}_{t,u}\boldsymbol{\upsilon} \\ \text{vec}(\mathbf{Q}_t) &= \mathbf{f}_{t,q} + \mathbf{D}_{t,q}\mathbf{q} \\ \text{vec}(\mathbf{Z}_t) &= \mathbf{f}_{t,z} + \mathbf{D}_{t,z}\boldsymbol{\zeta} \\ \text{vec}(\mathbf{a}_t) &= \mathbf{f}_{t,a} + \mathbf{D}_{t,a}\boldsymbol{\alpha} \\ \text{vec}(\mathbf{R}_t) &= \mathbf{f}_{t,r} + \mathbf{D}_{t,r}\mathbf{r} \\ \text{vec}(\boldsymbol{\Lambda}) &= \mathbf{f}_\lambda + \mathbf{D}_\lambda\boldsymbol{\lambda} \\ \text{vec}(\boldsymbol{\xi}) &= \mathbf{f}_\xi + \mathbf{D}_\xi\mathbf{p}\end{aligned} \tag{77}$$

The estimated parameters are now the column vectors, $\boldsymbol{\beta}$, $\boldsymbol{\upsilon}$, $\mathbf{q}$, $\boldsymbol{\zeta}$, $\boldsymbol{\alpha}$, $\mathbf{r}$, $\mathbf{p}$ and $\boldsymbol{\lambda}$. The time-varying aspect comes from the time-varying $\mathbf{f}$ and $\mathbf{D}$. Note that variance-covariance matrices must be positive-definite and we cannot specify a form that cannot be estimated. Fixing the diagonal terms and estimating the

[12] For example, if both $\mathbf{G}$ and $\mathbf{H}$ are column vectors, then the system is over-constrained and has no solution.

off-diagonals would not be allowed. Thus the $\mathbf{f}$ and $\mathbf{D}$ terms for $\mathbf{Q}$, $\mathbf{R}$ and $\mathbf{\Lambda}$ are limited. For the other parameters, the forms are fairly unrestricted, except that the $\mathbf{D}$s need to be full rank so that we are not specifying an under-constrained model. 'Full rank' will imply that we are not trying to estimate confounded matrix elements; for example, trying to estimate a_1 and a_2 but only $a_1 + a_2$ appear in the model.

The temporally variable MARSS model, equation 76 together with equation 77, looks rather different than other temporally variable MARSS models, such as a VARSSX or MARSS with covariates model, in the literature. But those models are special cases of this equation. By deriving an EM algorithm for this more general (if unfamiliar) form, I then have an algorithm for many different types of time-varying MARSS models with linear constraints on the parameter elements. Below I show some examples.

4.1 MARSS model with linear constraints

We can use equation 76 to put linear constraints on the elements of the parameters, $\mathbf{B}$, $\mathbf{u}$, $\mathbf{Q}$, $\mathbf{Z}$, $\mathbf{a}$, $\mathbf{R}$, $\boldsymbol{\xi}$ and $\mathbf{\Lambda}$. Here is an example of a simple MARSS model with linear constraints:

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

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

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

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

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

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

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

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

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

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

Table 3: Kronecker and vec relations. Here $\mathbf{A}$ is $n \times m$, $\mathbf{B}$ is $m \times p$, $\mathbf{C}$ is $p \times q$, and $\mathbf{E}$ and $\mathbf{D}$ are $p \times p$. $\mathbf{a}$ 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. (78)
$$\mathbf{a} = (\mathbf{a}^\top \otimes \mathbf{I}_1)$$

$$\text{vec}(\mathbf{Aa}) = (\mathbf{a}^\top \otimes \mathbf{I}_n)\,\text{vec}(\mathbf{A}) = \mathbf{Aa}$$
$$\text{vec}(\mathbf{Aa}) = \mathbf{Aa} \text{ since } \mathbf{Aa} \text{ is itself an } m \times 1 \text{ column vector.} \quad (79)$$

$$\text{vec}(\mathbf{AB}) = (\mathbf{I}_p \otimes \mathbf{A})\,\text{vec}(\mathbf{B}) = (\mathbf{B}^\top \otimes \mathbf{I}_n)\,\text{vec}(\mathbf{A}) \quad (80)$$

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\,\text{vec}(\mathbf{B}) \quad (81)$$

$$\text{vec}(\mathbf{a}^\top\mathbf{Ba}) = \mathbf{a}^\top\mathbf{Ba} = (\mathbf{a}^\top \otimes \mathbf{a})\,\text{vec}(\mathbf{B}) \quad (82)$$

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC} \otimes \mathbf{BD})$$
$$(\mathbf{A} \otimes \mathbf{B}) + (\mathbf{A} \otimes \mathbf{C}) = (\mathbf{A} \otimes (\mathbf{B} + \mathbf{C})) \quad (83)$$

$$(\mathbf{a} \otimes \mathbf{I}_p)\mathbf{C} = (\mathbf{a} \otimes \mathbf{C})$$
$$\mathbf{C}(\mathbf{a}^\top \otimes \mathbf{I}_q) = (\mathbf{a}^\top \otimes \mathbf{C}) \quad (84)$$
$$\mathbf{E}(\mathbf{a}^\top \otimes \mathbf{D}) = \mathbf{ED}(\mathbf{a}^\top \otimes \mathbf{I}_p) = (\mathbf{a}^\top \otimes \mathbf{ED})$$

$$(\mathbf{a} \otimes \mathbf{I}_p)\mathbf{C}(\mathbf{b}^\top \otimes \mathbf{I}_q) = (\mathbf{ab}^\top \otimes \mathbf{C}) \quad (85)$$

$$(\mathbf{a} \otimes \mathbf{b}) = \text{vec}(\mathbf{ba}^\top)$$
$$(\mathbf{a}^\top \otimes \mathbf{b}^\top) = (\mathbf{a} \otimes \mathbf{b})^\top = (\,\text{vec}(\mathbf{ba}^\top))^\top \quad (86)$$

$$(\mathbf{A}^\top \otimes \mathbf{B}^\top) = (\mathbf{A} \otimes \mathbf{B})^\top \quad (87)$$

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

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

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

4.2 A MARSS model with exogenous variables

The following is a commonly seen MARSS model with covariates $\mathbf{c}_t$ and $\mathbf{d}_t$ appearing as additive elements:

$$\begin{aligned}\boldsymbol{x}_t &= \mathbf{B}\boldsymbol{x}_{t-1} + \mathbf{C}\mathbf{c}_t + \boldsymbol{w}_t \\ \boldsymbol{y}_t &= \mathbf{Z}\boldsymbol{x}_t + \mathbf{D}\mathbf{d}_t + \boldsymbol{v}_t\end{aligned}$$

Here, $\mathbf{D}$ is the effect of $\mathbf{d}_t$ on $\boldsymbol{y}_t$ not a design matrix (which would have a subscript). We would typically want to estimate $\mathbf{C}$ or $\mathbf{D}$ which are the influence of our covariates on our responses, $\boldsymbol{x}$ or $\boldsymbol{y}$. Let's say there are p covariates in $\mathbf{c}_t$ and q covariates in $\mathbf{d}_t$. Then we can write the above in vec form:

$$\begin{aligned}\boldsymbol{x}_t &= (\boldsymbol{x}_{t-1}^\top \otimes \mathbf{I}_m)\,\mathrm{vec}(\mathbf{B}) + (\mathbf{c}_t^\top \otimes \mathbf{I}_p)\,\mathrm{vec}(\mathbf{C}) + \boldsymbol{w}_t \\ \boldsymbol{y}_t &= (\boldsymbol{x}_t^\top \otimes \mathbf{I}_n)\,\mathrm{vec}(\mathbf{Z}) + (\mathbf{d}_t^\top \otimes \mathbf{I}_q)\,\mathrm{vec}(\mathbf{D}) + \boldsymbol{v}_t\end{aligned} \tag{88}$$

Let's say we put no constraints $\mathbf{B}$, $\mathbf{Z}$, $\mathbf{Q}$, $\mathbf{R}$, $\boldsymbol{\xi}$, or $\boldsymbol{\Lambda}$. Then in the form of equation 76,

$$\begin{aligned}\boldsymbol{x}_t &= (\boldsymbol{x}_{t-1}^\top \otimes \mathbf{I}_m)\,\mathrm{vec}(\mathbf{B}_t) + \mathrm{vec}(\mathbf{u}_t) + \boldsymbol{w}_t \\ \boldsymbol{y}_t &= (\boldsymbol{x}_t^\top \otimes \mathbf{I}_n)\,\mathrm{vec}(\mathbf{Z}_t) + \mathrm{vec}(\mathbf{a}_t) + \boldsymbol{v}_t,\end{aligned}$$

with the parameters defined as follows:

$$\begin{aligned}\mathrm{vec}(\mathbf{B}_t) &= \mathbf{f}_{t,b} + \mathbf{D}_{t,b}\boldsymbol{\beta};\ \mathbf{f}_{t,b} = 0;\ \mathbf{D}_{t,b} = 1;\ \boldsymbol{\beta} = \mathrm{vec}(\mathbf{B}) \\ \mathrm{vec}(\mathbf{u}_t) &= \mathbf{f}_{t,u} + \mathbf{D}_{t,u}\boldsymbol{v};\ \mathbf{f}_{t,u} = 0;\ \mathbf{D}_{t,u} = (\mathbf{c}_t^\top \otimes \mathbf{I}_p);\ \boldsymbol{v} = \mathrm{vec}(\mathbf{C}) \\ \mathrm{vec}(\mathbf{Q}_t) &= \mathbf{f}_{t,q} + \mathbf{D}_{t,q}\mathbf{q};\ \mathbf{f}_{t,q} = 0;\ \mathbf{D}_{t,q} = \mathbf{D}_q \\ \mathrm{vec}(\mathbf{Z}_t) &= \mathbf{f}_{t,z} + \mathbf{D}_{t,z}\boldsymbol{\zeta};\ \mathbf{f}_{t,z} = 0;\ \mathbf{D}_{t,z} = 1;\ \boldsymbol{\zeta} = \mathrm{vec}(\mathbf{Z}) \\ \mathrm{vec}(\mathbf{a}_t) &= \mathbf{f}_{t,a} + \mathbf{D}_{t,a}\boldsymbol{\alpha};\ \mathbf{f}_{t,a} = 0;\ \mathbf{D}_{t,a} = (\mathbf{d}_t^\top \otimes \mathbf{I}_q);\ \boldsymbol{\alpha} = \mathrm{vec}(\mathbf{D}) \\ \mathrm{vec}(\mathbf{R}_t) &= \mathbf{f}_{t,r} + \mathbf{D}_{t,r}\mathbf{r};\ \mathbf{f}_{t,r} = 0;\ \mathbf{D}_{t,r} = \mathbf{D}_r \\ \mathrm{vec}(\boldsymbol{\Lambda}) &= \mathbf{f}_\lambda + \mathbf{D}_\lambda\boldsymbol{\lambda};\ \mathbf{f}_\lambda = 0 \\ \mathrm{vec}(\boldsymbol{\xi}) &= \boldsymbol{\xi} = \mathbf{f}_\xi + \mathbf{D}_\xi\mathbf{p};\ \mathbf{f}_\xi = 0;\ \mathbf{D}_\xi = 1\end{aligned}$$

Note that variance-covariance matrices are never unconstrained really so we use $\mathbf{D}_q$, $\mathbf{D}_r$ and $\mathbf{D}_\lambda$ to specify the symmetry within the matrix.

The transformation of the simple MARSS with covariates (equation 88) into the form of equation 76 may seem a little painful, but the advantage is that a single EM algorithm can be used for a large class of models. Presumably, the transformation of the equation will be hidden from users by a wrapper function that does the reformulation before passing the model to the general EM algorithm. In the MARSS R package, this reformulation is done in the `MARSS.marxss` function.

4.3 A general MARSS model with exogenous variables

Let's imagine now a very general MARSS model with various 'inputs'. ' input' here just means that it is some fully known matrix rather than something we are estimating. It could be a sequence of 0s and 1s if for example we were fitting a before/after sort of model. Below the letters with a t subscript are the inputs (and $\mathbf{D}_t$ is an input not a design matrix), except $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{w}$ and $\boldsymbol{v}$.

$$\begin{aligned}\boldsymbol{x}_t &= \mathbf{J}_t\mathbf{B}\mathbf{L}_t\boldsymbol{x}_{t-1} + \mathbf{C}_t\mathbf{U}\mathbf{c}_t + \mathbf{G}_t\boldsymbol{w}_t \\ \boldsymbol{y}_t &= \mathbf{M}_t\mathbf{Z}\mathbf{N}_t\boldsymbol{x}_t + \mathbf{D}_t\mathbf{A}\mathbf{d}_t + \mathbf{H}_t\boldsymbol{v}_t\end{aligned} \tag{89}$$

In vec form, this is:

$$\begin{aligned}
\boldsymbol{x}_t &= (\boldsymbol{x}_{t-1}^\top \otimes \mathbf{I}_m)(\mathbf{L}_t^\top \otimes \mathbf{J}_t)\,\mathrm{vec}(\mathbf{B}) + (\mathbf{c}_t^\top \otimes \mathbf{C}_t)\,\mathrm{vec}(\mathbf{U}) + \mathbf{G}_t\boldsymbol{w}_t \\
&= (\boldsymbol{x}_{t-1}^\top \otimes \mathbf{I}_m)(\mathbf{L}_t^\top \otimes \mathbf{J}_t)(\mathbf{f}_b + \mathbf{D}_b\boldsymbol{\beta}) + (\mathbf{c}_t^\top \otimes \mathbf{C}_t)(\mathbf{f}_u + \mathbf{D}_u\boldsymbol{v}) + \mathbf{G}_t\boldsymbol{w}_t \\
\boldsymbol{W}_t &\sim \mathrm{MVN}(0, \mathbf{G}_t\mathbf{Q}\mathbf{G}_t^\top)
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{y}_t &= (\boldsymbol{x}_t^\top \otimes \mathbf{I}_n)(\mathbf{N}_t^\top \otimes \mathbf{M}_t)\,\mathrm{vec}(\mathbf{Z}) + (\mathbf{d}_t^\top \otimes \mathbf{D}_t)\,\mathrm{vec}(\mathbf{A}) + \mathbf{H}_t\boldsymbol{v}_t \\
&= (\boldsymbol{x}_t^\top \otimes \mathbf{I}_n)\mathbb{Z}_t(\mathbf{f}_z + \mathbf{D}_z\boldsymbol{\zeta}) + \mathbb{A}_t(\mathbf{f}_a + \mathbf{D}_a\boldsymbol{\alpha}) + \mathbf{H}_t\boldsymbol{v}_t \\
\boldsymbol{V}_t &\sim \mathrm{MVN}(0, \mathbf{H}_t\mathbf{R}\mathbf{H}_t^\top)
\end{aligned} \tag{90}$$

$$\boldsymbol{X}_{t_0} \sim \mathrm{MVN}(\mathbf{f}_\xi + \mathbf{D}_\xi\mathbf{p}, \mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top), \text{ where } \mathrm{vec}(\boldsymbol{\Lambda}) = \mathbf{f}_\lambda + \mathbf{D}_\lambda\boldsymbol{\lambda}$$

We could write down a likelihood function for this model but written this way, the model presumes that $\mathbf{H}_t\mathbf{R}\mathbf{H}_t^\top$, $\mathbf{G}_t\mathbf{Q}\mathbf{G}_t^\top$, and $\mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top$ are valid variance-covariance matrices. I will actually write this model differently below because I don't want to make that assumption.

We define the $\mathbf{f}$ and $\mathbf{D}$ parameters as follows.

$$\begin{aligned}
\mathrm{vec}(\mathbf{B}_t) &= \mathbf{f}_{t,b} + \mathbf{D}_{t,b}\boldsymbol{\beta} = (\mathbf{L}_t^\top \otimes \mathbf{J}_t)\mathbf{f}_b + (\mathbf{L}_t^\top \otimes \mathbf{J}_t)\mathbf{D}_b\boldsymbol{\beta} \\
\mathrm{vec}(\mathbf{u}_t) &= \mathbf{f}_{t,u} + \mathbf{D}_{t,u}\boldsymbol{v} = (\mathbf{c}_t^\top \otimes \mathbf{C}_t)\mathbf{f}_u + (\mathbf{c}_t^\top \otimes \mathbf{C}_t)\mathbf{D}_u\boldsymbol{v} \\
\mathrm{vec}(\mathbf{Q}_t) &= \mathbf{f}_{t,q} + \mathbf{D}_{t,q}\mathbf{q} = (\mathbf{G}_t \otimes \mathbf{G}_t)\mathbf{f}_q + (\mathbf{G}_t \otimes \mathbf{G}_t)\mathbf{D}_q\mathbf{q} \\
\mathrm{vec}(\mathbf{Z}_t) &= \mathbf{f}_{t,z} + \mathbf{D}_{t,z}\boldsymbol{\zeta} = (\mathbf{N}_t^\top \otimes \mathbf{M}_t)\mathbf{f}_z + (\mathbf{N}_t^\top \otimes \mathbf{M}_t)\mathbf{D}_z\boldsymbol{\zeta} \\
\mathrm{vec}(\mathbf{a}_t) &= \mathbf{f}_{t,a} + \mathbf{D}_{t,a}\boldsymbol{\alpha} = (\mathbf{d}_t^\top \otimes \mathbf{D}_t)\mathbf{f}_a + (\mathbf{d}_t^\top \otimes \mathbf{D}_t)\mathbf{D}_a\boldsymbol{\alpha} \\
\mathrm{vec}(\mathbf{R}_t) &= \mathbf{f}_{t,r} + \mathbf{D}_{t,r}\mathbf{r} = (\mathbf{H}_t \otimes \mathbf{H}_t)\mathbf{f}_q + (\mathbf{H}_t \otimes \mathbf{H}_t)\mathbf{D}_r\mathbf{r} \\
\mathrm{vec}(\boldsymbol{\Lambda}) &= \mathbf{f}_\lambda + \mathbf{D}_\lambda\boldsymbol{\lambda} = 0 + \mathbf{D}_\lambda\boldsymbol{\lambda} \\
\mathrm{vec}(\boldsymbol{\xi}) &= \boldsymbol{\xi} = \mathbf{f}_\xi + \mathbf{D}_\xi\mathbf{p} = 0 + 1\mathbf{p}
\end{aligned}$$

Here, for example $\mathbf{f}_b$ and $\mathbf{D}_b$ indicate the linear constraints on $\mathbf{B}$ and $\mathbf{f}_{t,b}$ is $(\mathbf{L}_t^\top \otimes \mathbf{J}_t)\mathbf{f}_b$ and $\mathbf{D}_{t,b}$ is $(\mathbf{L}_t^\top \otimes \mathbf{J}_t)\mathbf{D}_b$. The elements of $\mathbf{B}$ that are being estimated are $\boldsymbol{\beta}$ arranged as a column vector.

As usual, this reformulation looks cumbersome, but would be hidden from the user presumably.

4.4 The expected log-likelihood function

As mentioned above, we do not necessarily want to assume that $\mathbf{H}_t\mathbf{R}_t\mathbf{H}_t^\top$, $\mathbf{G}_t\mathbf{Q}_t\mathbf{G}_t^\top$, and $\mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top$ are valid variance-covariance matrices. This would rule out many MARSS models that we would like to fit. For example, if $\mathbf{Q} = \sigma^2$ and $\mathbf{G} = \begin{bmatrix}1\\1\\1\end{bmatrix}$, $\mathbf{G}\mathbf{Q}\mathbf{G}^\top$ would be an invalid variance-variance matrix. However, this is a valid MARSS model. We do need to be careful that $\mathbf{H}_t$ and $\mathbf{G}_t$ are specified such that the model has a solution. For example, a model where both $\mathbf{G}$ and $\mathbf{H}$ are $\begin{bmatrix}1\\1\\1\end{bmatrix}$ would not be solvable for all $\boldsymbol{y}$.

Instead I will define $\Phi_t = (\mathbf{G}_t^\top\mathbf{G}_t)^{-1}\mathbf{G}_t^\top$, $\Xi_t = (\mathbf{H}_t^\top\mathbf{H}_t)^{-1}\mathbf{H}_t^\top$, and $\Pi = (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top$. I then require that the inverses of $\mathbf{G}_t^\top\mathbf{G}_t$, $\mathbf{H}_t^\top\mathbf{H}_t$, and $\mathbf{F}^\top\mathbf{F}$ exist and that $\mathbf{f}_{t,q} + \mathbf{D}_{t,q}\mathbf{q}$, $\mathbf{f}_{t,r} + \mathbf{D}_{t,r}\mathbf{r}$, and $\mathbf{f}_\lambda + \mathbf{D}_\lambda\boldsymbol{\lambda}$ specify valid variance-covariance matrices. These are much less stringent restrictions.

For the purpose of writing down the expected log-likelihood, our MARSS model is now written

$$\begin{aligned}
\Phi_t\boldsymbol{x}_t &= \Phi_t(\boldsymbol{x}_{t-1}^\top \otimes \mathbf{I}_m)\,\mathrm{vec}(\mathbf{B}_t) + \Phi_t\,\mathrm{vec}(\mathbf{u}_t) + \boldsymbol{w}_t, \quad \text{where } \boldsymbol{W}_t \sim \mathrm{MVN}(0, \mathbf{Q}_t) \\
\Xi_t\boldsymbol{y}_t &= \Xi_t(\boldsymbol{x}_t^\top \otimes \mathbf{I}_n)\,\mathrm{vec}(\mathbf{Z}_t) + \Xi_t\,\mathrm{vec}(\mathbf{a}_t) + \boldsymbol{v}_t, \quad \text{where } \boldsymbol{V}_t \sim \mathrm{MVN}(0, \mathbf{R}_t) \\
\Pi\boldsymbol{x}_{t_0} &= \Pi\boldsymbol{\xi} + \boldsymbol{l}, \quad \text{where } \boldsymbol{L} \sim \mathrm{MVN}(0, \boldsymbol{\Lambda})
\end{aligned} \tag{91}$$

As mentioned before, this relies on $\mathbf{G}$ and $\mathbf{H}$ having forms that do not lead to over- or under-constrained linear systems.

To derive the EM update equations, we need the expected log-likelihood function for the time-varying MARSS model. Using equation 91, we get

$$\begin{aligned}
\mathrm{E}_{\mathbf{XY}}[\log \boldsymbol{L}(\boldsymbol{Y},\boldsymbol{X};\Theta)] = -\frac{1}{2}\mathrm{E}_{\mathbf{XY}}\Bigg(& \sum_1^T (\boldsymbol{Y}_t - (\boldsymbol{X}_t^\top \otimes \mathbf{I}_m)\,\mathrm{vec}(\mathbf{Z}_t) - \mathrm{vec}(\mathbf{a}_t))^\top \Xi_t^\top \mathbf{R}_t^{-1} \Xi_t \\
& (\boldsymbol{Y}_t - (\boldsymbol{X}_t^\top \otimes \mathbf{I}_m)\,\mathrm{vec}(\mathbf{Z}_t) - \mathrm{vec}(\mathbf{a}_t)) + \sum_1^T \log|\mathbf{R}_t| \\
& + \sum_{t_0+1}^T (\boldsymbol{X}_t - (\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I}_m)\,\mathrm{vec}(\mathbf{B}_t) - \mathrm{vec}(\mathbf{u}_t))^\top \Phi_t^\top \mathbf{Q}_t^{-1} \Phi_t \\
& (\boldsymbol{X}_t - (\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I}_m)\,\mathrm{vec}(\mathbf{B}_t) - \mathrm{vec}(\mathbf{u}_t)) + \sum_{t_0+1}^T \log|\mathbf{Q}_t| \\
& + (\boldsymbol{X}_{t_0} - \mathrm{vec}(\boldsymbol{\xi}))^\top \Pi^\top \boldsymbol{\Lambda}^{-1} \Pi (\boldsymbol{X}_{t_0} - \mathrm{vec}(\boldsymbol{\xi})) + \log|\boldsymbol{\Lambda}| + \log 2\pi \Bigg)
\end{aligned} \tag{92}$$

If any $\mathbf{G}_t$, $\mathbf{H}_t$ or $\mathbf{F}$ is all zero, then the line in the likelihood with $\mathbf{R}_t$, $\mathbf{Q}_t$ or $\boldsymbol{\Lambda}$, respectively, does not appear. If any $\boldsymbol{x}_{t_0}$ are fixed, meaning all zero row in $\mathbf{F}$, that $\boldsymbol{X}_{t_0} \equiv \boldsymbol{\xi}$ anywhere it appears in the likelihood. The way I have written the general equation, some $\boldsymbol{x}_{t_0}$ might be fixed and others stochastic.

The vec of the model parameters are defined as follows:

$$\begin{aligned}
\mathrm{vec}(\mathbf{B}_t) &= \mathbf{f}_{t,b} + \mathbf{D}_{t,b}\boldsymbol{\beta} \\
\mathrm{vec}(\mathbf{u}_t) &= \mathbf{f}_{t,u} + \mathbf{D}_{t,u}\boldsymbol{\upsilon} \\
\mathrm{vec}(\mathbf{Z}_t) &= \mathbf{f}_{t,z} + \mathbf{D}_{t,z}\boldsymbol{\zeta} \\
\mathrm{vec}(\mathbf{a}_t) &= \mathbf{f}_{t,a} + \mathbf{D}_{t,a}\boldsymbol{\alpha} \\
\mathrm{vec}(\mathbf{Q}_t) &= \mathbf{f}_{t,q} + \mathbf{D}_{t,q}\mathbf{q} \\
\mathrm{vec}(\mathbf{R}_t) &= \mathbf{f}_{t,r} + \mathbf{D}_{t,r}\mathbf{r} \\
\mathrm{vec}(\boldsymbol{\xi}) &= \mathbf{f}_{\xi} + \mathbf{D}_{\xi}\mathbf{p} \\
\mathrm{vec}(\boldsymbol{\Lambda}) &= \mathbf{f}_{\lambda} + \mathbf{D}_{\lambda}\boldsymbol{\lambda} \\
\Phi_t &= (\mathbf{G}_t^\top \mathbf{G}_t)^{-1}\mathbf{G}_t^\top \\
\Xi_t &= (\mathbf{H}_t^\top \mathbf{H}_t)^{-1}\mathbf{H}_t^\top \\
\Pi &= (\mathbf{F}^\top \mathbf{F})^{-1}\mathbf{F}^\top
\end{aligned}$$

5 The constrained update equations

The derivation proceeds by taking the partial derivative of equation 92 with respect to the estimated terms, the $\boldsymbol{\zeta}$, $\boldsymbol{\alpha}$, etc, setting the derivative to zero, and solving for those estimated terms. Conceptually, the algebraic steps in the derivation are similar to those in the unconstrained derivation. See the notes in Sections 3 and 2 regarding implementation of the EM algorithm when Θ is broken into parts (e.g., $\mathbf{B}$, $\mathbf{u}$, $\mathbf{Q}$, etc.).

5.1 The general u update equations

We take the derivative of Ψ (equation 92) with respect to $\boldsymbol{\upsilon}$.

$$\begin{aligned}
\partial\Psi/\partial\boldsymbol{\upsilon} = -\frac{1}{2}\sum_{t=1}^T \Bigg(& -\partial(\,\mathrm{E}[\boldsymbol{X}_t^\top \mathbb{Q}_t \mathbf{D}_{t,u}\boldsymbol{\upsilon}])/\partial\boldsymbol{\upsilon} - \partial(\,\mathrm{E}[\boldsymbol{\upsilon}^\top \mathbf{D}_{t,u}^\top \mathbb{Q}_t \boldsymbol{X}_t])/\partial\boldsymbol{\upsilon} \\
& + \partial(\,\mathrm{E}[((\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I}_m)\,\mathrm{vec}(\mathbf{B}_t))^\top \mathbb{Q}_t \mathbf{D}_{t,u}\boldsymbol{\upsilon}])/\partial\boldsymbol{\upsilon} + \partial(\,\mathrm{E}[\boldsymbol{\upsilon}^\top \mathbf{D}_{t,u}^\top \mathbb{Q}_t (\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I}_m)\,\mathrm{vec}(\mathbf{B}_t)])/\partial\boldsymbol{\upsilon} \\
& + \partial(\boldsymbol{\upsilon}^\top \mathbf{D}_{t,u}^\top \mathbb{Q}_t \mathbf{D}_{t,u}\boldsymbol{\upsilon})/\partial\boldsymbol{\upsilon} + \partial(\,\mathrm{E}[\mathbf{f}_{t,u}^\top \mathbb{Q}_t \mathbf{D}_{t,u}\boldsymbol{\upsilon}])/\partial\boldsymbol{\upsilon} + \partial(\,\mathrm{E}[\boldsymbol{\upsilon}^\top \mathbf{D}_{t,u}^\top \mathbb{Q}_t \mathbf{f}_{t,u}])/\partial\boldsymbol{\upsilon} \Bigg)
\end{aligned} \tag{93}$$

where $\mathbb{Q}_t = \Phi_t^\top \mathbf{Q}_t^{-1} \Phi_t$.

Since $\boldsymbol{v}$ is to the far left or right in each term, the derivative is simple using the derivative terms in table 3.1. $\partial\Psi/\partial\boldsymbol{v}$ becomes:

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{v} = -\frac{1}{2}\sum_{t=1}^{T}\Big(&- 2\,\mathrm{E}[\boldsymbol{X}_t^\top\mathbb{Q}_t\mathbf{D}_{t,u}] + 2\,\mathrm{E}[((\boldsymbol{X}_{t-1}^\top\otimes\mathbf{I}_m)\,\mathrm{vec}(\mathbf{B}_t))^\top\mathbb{Q}_t\mathbf{D}_{t,u}] \\ &+ 2(\boldsymbol{v}^\top\mathbf{D}_{t,u}^\top\mathbb{Q}_t\mathbf{D}_{t,u}) + 2\,\mathrm{E}[\mathbf{f}_{t,u}^\top\mathbb{Q}_t\mathbf{D}_{t,u}]\Big)\end{aligned} \tag{94}$$

Set the left side to zero and transpose the whole equation.

$$\mathbf{0} = \sum_{t=1}^{T}\Big(\mathbf{D}_{t,u}^\top\mathbb{Q}_t\,\mathrm{E}[\boldsymbol{X}_t] - \mathbf{D}_{t,u}^\top\mathbb{Q}_t(\,\mathrm{E}[\boldsymbol{X}_{t-1}]^\top\otimes\mathbf{I}_m)\,\mathrm{vec}(\mathbf{B}_t) - \mathbf{D}_{t,u}^\top\mathbb{Q}_t\mathbf{D}_{t,u}\boldsymbol{v} - \mathbf{D}_{t,u}^\top\mathbb{Q}_t\mathbf{f}_{t,u}\Big) \tag{95}$$

Thus,

$$(\sum_{t=1}^{T}\mathbf{D}_{t,u}^\top\mathbb{Q}_t\mathbf{D}_{t,u})\boldsymbol{v} = \sum_{t=1}^{T}\mathbf{D}_{t,u}^\top\mathbb{Q}_t\big(\,\mathrm{E}[\boldsymbol{X}_t] - (\,\mathrm{E}[\boldsymbol{X}_{t-1}]^\top\otimes\mathbf{I}_m)\,\mathrm{vec}(\mathbf{B}_t) - \mathbf{f}_{t,u}\big) \tag{96}$$

We solve for $\boldsymbol{v}$, and the new $\boldsymbol{v}$ for the $j+1$ iteration of the EM algorithm is

$$\boldsymbol{v}_{j+1} = \Big(\sum_{t=1}^{T}\mathbf{D}_{t,u}^\top\mathbb{Q}_t\mathbf{D}_{t,u}\Big)^{-1}\sum_{t=1}^{T}\mathbf{D}_{t,u}^\top\mathbb{Q}_t\big(\widetilde{\mathbf{x}}_t - (\widetilde{\mathbf{x}}_{t-1}^\top\otimes\mathbf{I}_m)\,\mathrm{vec}(\mathbf{B}_t) - \mathbf{f}_{t,u}\big) \tag{97}$$

where $\mathbb{Q}_t = \Phi_t^\top\mathbf{Q}_t^{-1}\Phi_t = \mathbf{G}_t(\mathbf{G}_t^\top\mathbf{G}_t)^{-1}\mathbf{Q}_t^{-1}(\mathbf{G}_t^\top\mathbf{G}_t)^{-1}\mathbf{G}_t^\top$.

The update equation requires that $\sum_{t=1}^{T}\mathbf{D}_{t,u}^\top\mathbb{Q}_t\mathbf{D}_{t,u}$ is invertible. It generally will be if $\Phi_t\mathbf{Q}_t\Phi_t^\top$ is a proper variance-covariance matrix (positive semi-definite) and $\mathbf{D}_{t,u}$ is full rank. If $\mathbf{G}_t$ has all-zero rows then $\Phi_t\mathbf{Q}_t\Phi_t^\top$ has zeros on the diagonal and we have a partially deterministic model. In this case, $\mathbb{Q}_t$ will have all-zero row/columns and $\mathbf{D}_{t,u}^\top\mathbb{Q}_t\mathbf{D}_{t,u}$ will not be invertible unless the corresponding row of $\mathbf{D}_{t,u}$ is zero. This means that if one of the $\boldsymbol{x}$ rows is fully deterministic then the corresponding row of $\mathbf{u}$ would need to be fixed. We can get around this, however. See section 7 on the modifications to the update equation when some of the $\boldsymbol{x}$'s are fully deterministic.

5.2 The general a update equation

The derivation of the update equation for $\boldsymbol{\alpha}$ with fixed and shared values is completely analogous to the derivation for $\boldsymbol{v}$. We take the derivative of Ψ with respect to $\boldsymbol{\alpha}$ and arrive at the analogous:

$$\begin{aligned}\boldsymbol{\alpha}_{j+1} &= \big(\sum_{t=1}^{T}\mathbf{D}_{t,a}^\top\mathbb{R}_t\mathbf{D}_{t,a}\big)^{-1}\sum_{t=1}^{T}\mathbf{D}_{t,a}^\top\mathbb{R}_t\big(\widetilde{\mathbf{y}}_t - (\widetilde{\mathbf{x}}_t^\top\otimes\mathbf{I}_n)\,\mathrm{vec}(\mathbf{Z}_t) - \mathbf{f}_{t,a}\big) \\ &= \big(\sum_{t=1}^{T}\mathbf{D}_{t,a}^\top\mathbb{R}_t\mathbf{D}_{t,a}\big)^{-1}\sum_{t=1}^{T}\mathbf{D}_{t,a}^\top\mathbb{R}_t\big(\widetilde{\mathbf{y}}_t - \mathbf{Z}_t\widetilde{\mathbf{x}}_t - \mathbf{f}_{t,a}\big)\end{aligned} \tag{98}$$

$\sum_{t=1}^{T}\mathbf{D}_{t,a}^\top\mathbb{R}_t\mathbf{D}_{t,a}$ must be invertible.

5.3 The general $\boldsymbol{\xi}$ update equation, stochastic initial state

When $\boldsymbol{x}_0$ is treated as stochastic with an unknown mean and known variance, the derivation of the update equation for $\boldsymbol{\xi}$ with fixed and shared values is as follows. Take the derivative of Ψ (using equation 92) with respect to $\mathbf{p}$:

$$\partial\Psi/\partial\mathbf{p} = \big(\widetilde{\mathbf{x}}_0^\top\mathbb{L} - \boldsymbol{\xi}^\top\mathbb{L}\big) \tag{99}$$

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

$$\mathbf{0} = \mathbf{D}_\xi^\top\big(\mathbb{L}\widetilde{\mathbf{x}}_0 - \mathbb{L}\mathbf{f}_\xi + \mathbb{L}\mathbf{D}_\xi\mathbf{p}\big) \tag{100}$$

Thus,

$$\mathbf{p}_{j+1} = \big(\mathbf{D}_\xi^\top\mathbb{L}\mathbf{D}_\xi\big)^{-1}\mathbf{D}_\xi^\top\mathbb{L}(\widetilde{\mathbf{x}}_0 - \mathbf{f}_\xi) \tag{101}$$

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

$$\boldsymbol{\xi}_{j+1} = \mathbf{f}_\xi + \mathbf{D}_\xi\mathbf{p}_{j+1}, \tag{102}$$

When the initial state is defined as at $t=1$, replace $\widetilde{\mathbf{x}}_0$ with $\widetilde{\mathbf{x}}_1$ in equation 101.

5.4 The general $\boldsymbol{\xi}$ update equation, fixed $\boldsymbol{x}_0$

For this case, $\boldsymbol{x}_0$ is treated as fixed, i.e., as another parameter, and $\boldsymbol{\Lambda}$ does not appear in the equation. It will be easier to work with Ψ written as follows:

$$\begin{aligned}\mathrm{E}_{\mathbf{XY}}[\log \boldsymbol{L}(\boldsymbol{Y},\boldsymbol{X};\Theta)] = -\frac{1}{2}\,\mathrm{E}_{\mathbf{XY}}\Bigg(&\sum_1^T(\boldsymbol{Y}_t-\mathbf{Z}_t\boldsymbol{X}_t-\mathbf{a}_t)^\top\mathbb{R}_t(\boldsymbol{Y}_t-\mathbf{Z}_t\boldsymbol{X}_t-\mathbf{a}_t)+\sum_1^T\log|\mathbf{R}_t|\\&+\sum_2^T(\boldsymbol{X}_t-\mathbf{B}_t\boldsymbol{X}_{t-1}-\mathbf{u}_t)^\top\mathbb{Q}_t(\boldsymbol{X}_t-\mathbf{B}_t\boldsymbol{X}_{t-1}-\mathbf{u}_t)+\sum_1^T\log|\mathbf{Q}_t|+\log 2\pi\Bigg)\\&\boldsymbol{x}_0\equiv\mathbf{f}_\xi+\mathbf{D}_\xi\mathbf{p}\end{aligned}\tag{103}$$

This is the same as equation 92 except not written in vec form and $\boldsymbol{\Lambda}$ does not appear. Take the derivative of Ψ using equation 103. Terms not involving $\mathbf{p}$ will drop out:

$$\begin{aligned}\partial\Psi/\partial\mathbf{p} = -\frac{1}{2}\Big(&-\mathrm{E}[\partial(\mathbb{P}_1^\top\mathbb{Q}_1\mathbf{B}_1\mathbf{D}_\xi\mathbf{p})/\partial\mathbf{p}]-\mathrm{E}[\partial(\mathbf{p}^\top(\mathbf{B}_1\mathbf{D}_\xi)^\top\mathbb{Q}_1\mathbb{P}_1)/\partial\mathbf{p}]\\&+\mathrm{E}[\partial(\mathbf{p}^\top(\mathbf{B}_1\mathbf{D}_\xi)^\top\mathbb{Q}_1\mathbf{B}_1\mathbf{D}_\xi\mathbf{p})/\partial\mathbf{p}]\Big)\end{aligned}\tag{104}$$

where

$$\mathbb{P}_1 = \boldsymbol{X}_1-\mathbf{B}_1\mathbf{f}_\xi-\mathbf{u}_1\tag{105}$$

After pulling the constants out of the expectations and taking the derivative, we arrive at:

$$\partial\Psi/\partial\mathbf{p} = -\frac{1}{2}\Big(-2\,\mathrm{E}[\mathbb{P}_1]^\top\mathbb{Q}_1\mathbf{B}_1\mathbf{D}_\xi+2\mathbf{p}^\top(\mathbf{B}_1\mathbf{D}_\xi)^\top\mathbb{Q}_1\mathbf{B}_1\mathbf{D}_\xi\Big)\tag{106}$$

Set the left side to zero, and solve for $\mathbf{p}$.

$$\mathbf{p} = (\mathbf{D}_\xi^\top\mathbf{B}_1^\top\mathbb{Q}_1\mathbf{B}_1\mathbf{D}_\xi)^{-1}\mathbf{D}_\xi^\top\mathbf{B}_1^\top\mathbb{Q}_1(\widetilde{\mathbf{x}}_1-\mathbf{B}_1\mathbf{f}_\xi-\mathbf{u}_1)\tag{107}$$

This equation requires that the inverse right of the = exists and it might not if $\mathbf{B}_t$ or $\mathbb{Q}_1$ has any all zero rows/columns. In that case, defining $\boldsymbol{\xi}\equiv\boldsymbol{x}_1$ might work (section 5.5) or the problematic rows of $\boldsymbol{\xi}$ could be fixed. The new $\boldsymbol{\xi}$ is then,

$$\boldsymbol{\xi}_{j+1} = \mathbf{f}_\xi+\mathbf{D}_\xi\mathbf{p}_{j+1},\tag{108}$$

5.5 The general $\boldsymbol{\xi}$ update equation, fixed $\boldsymbol{x}_1$

When $\boldsymbol{x}_1$ is treated as fixed, i.e., as an estimated parameter, and $\boldsymbol{\Lambda}$ does not appear, the expected log likelihood, Ψ, is written as follows:

$$\begin{aligned}\mathrm{E}_{\mathbf{XY}}[\log \boldsymbol{L}(\boldsymbol{Y},\boldsymbol{X};\Theta)] = -\frac{1}{2}\,\mathrm{E}_{\mathbf{XY}}\Bigg(&\sum_1^T(\boldsymbol{Y}_t-\mathbf{Z}_t\boldsymbol{X}_t-\mathbf{a}_t)^\top\mathbb{R}_t(\boldsymbol{Y}_t-\mathbf{Z}_t\boldsymbol{X}_t-\mathbf{a}_t)+\sum_1^T\log|\mathbf{R}_t|\\&+\sum_2^T(\boldsymbol{X}_t-\mathbf{B}_t\boldsymbol{X}_{t-1}-\mathbf{u}_t)^\top\mathbb{Q}_t(\boldsymbol{X}_t-\mathbf{B}_t\boldsymbol{X}_{t-1}-\mathbf{u}_t)+\sum_2^T\log|\mathbf{Q}_t|+\log 2\pi\Bigg)\\&\boldsymbol{x}_1\equiv\mathbf{f}_\xi+\mathbf{D}_\xi\mathbf{p}\end{aligned}\tag{109}$$

Take the derivative of Ψ using equation 109:

$$\begin{aligned}\partial\Psi/\partial\mathbf{p} = -\frac{1}{2}\Big(&-\mathrm{E}[\partial(\mathbb{O}_1^\top\mathbb{R}_1\mathbf{Z}_1\mathbf{D}_\xi\mathbf{p})/\partial\mathbf{p}]-\mathrm{E}[\partial((\mathbf{Z}_1\mathbf{D}_\xi\mathbf{p})^\top\mathbb{R}_1\mathbb{O}_1)/\partial\mathbf{p}]\\&+\mathrm{E}[\partial((\mathbf{Z}_1\mathbf{D}_\xi\mathbf{p})^\top\mathbb{R}_1\mathbf{Z}_1\mathbf{D}_\xi\mathbf{p})/\partial\mathbf{p}]-\mathrm{E}[\partial(\mathbb{P}_2^\top\mathbb{Q}_2\mathbf{B}_2\mathbf{D}_\xi\mathbf{p})/\partial\mathbf{p}]-\mathrm{E}[\partial((\mathbf{B}_2\mathbf{D}_\xi\mathbf{p})^\top\mathbb{Q}_2\mathbb{P}_2)/\partial\mathbf{p}]\\&+\mathrm{E}[\partial((\mathbf{B}_2\mathbf{D}_\xi\mathbf{p})^\top\mathbb{Q}_2\mathbf{B}_2\mathbf{D}_\xi\mathbf{p})/\partial\mathbf{p}]\Big)\end{aligned}\tag{110}$$

where

$$\begin{aligned}\mathbb{P}_2 &= \boldsymbol{X}_2 - \mathbf{B}_2\mathbf{f}_\xi - \mathbf{u}_2 \\ \mathbb{O}_1 &= \boldsymbol{Y}_1 - \mathbf{Z}_1\mathbf{f}_\xi - \mathbf{a}_1\end{aligned} \tag{111}$$

In terms of the Kalman smoother output the new $\boldsymbol{\xi}$ for EM iteration $j+1$ when $\boldsymbol{\xi} \equiv \boldsymbol{x}_1$ is

$$\mathbf{p}_{j+1} = ((\mathbf{Z}_1\mathbf{D}_\xi)^\top \mathbb{R}_1 \mathbf{Z}_1\mathbf{D}_\xi + (\mathbf{B}_2\mathbf{D}_\xi)^\top \mathbb{Q}_2 \mathbf{B}_2\mathbf{D}_\xi)^{-1}((\mathbf{Z}_1\mathbf{D}_\xi)^\top \mathbb{R}_1 \widetilde{\mathbb{O}}_1 + (\mathbf{B}_2\mathbf{D}_\xi)^\top \mathbb{Q}_2 \widetilde{\mathbb{P}}_2) \tag{112}$$

where

$$\begin{aligned}\widetilde{\mathbb{P}}_2 &= \widetilde{\mathbf{x}}_2 - \mathbf{B}_2\mathbf{f}_\xi - \mathbf{u}_2 \\ \widetilde{\mathbb{O}}_1 &= \widetilde{\mathbf{y}}_1 - \mathbf{Z}_1\mathbf{f}_\xi - \mathbf{a}_1\end{aligned} \tag{113}$$

The new $\boldsymbol{\xi}$ is

$$\boldsymbol{\xi}_{j+1} = \mathbf{f}_\xi + \mathbf{D}_\xi \mathbf{p}_{j+1}, \tag{114}$$

5.6 The general B update equation

Take the derivative of Ψ with respect to $\boldsymbol{\beta}$; terms in Ψ do not involve $\boldsymbol{\beta}$ will equal 0 and drop out.

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\beta} = -\frac{1}{2}\sum_{t=1}^{T}\Big(&- \partial(\,\mathrm{E}[\boldsymbol{X}_t^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t \mathbf{D}_{t,b}\boldsymbol{\beta}])/\partial\boldsymbol{\beta} - \partial(\,\mathrm{E}[(\boldsymbol{\Upsilon}_t\mathbf{D}_{t,b}\boldsymbol{\beta})^\top \mathbb{Q}_t \boldsymbol{X}_t])/\partial\boldsymbol{\beta} \\ &+ \partial(\,\mathrm{E}[(\boldsymbol{\Upsilon}_t\mathbf{D}_{t,b}\boldsymbol{\beta})^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t \mathbf{D}_{t,b}\boldsymbol{\beta}])/\partial\boldsymbol{\beta} + \partial(\,\mathrm{E}[\mathbf{u}_t^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t \mathbf{D}_{t,b}\boldsymbol{\beta}])/\partial\boldsymbol{\beta} + \partial((\boldsymbol{\Upsilon}_t\mathbf{D}_{t,b}\boldsymbol{\beta})^\top \mathbb{Q}_t \mathbf{u}_t)/\partial\boldsymbol{\beta} \\ &+ \partial(\,\mathrm{E}[(\boldsymbol{\Upsilon}_t\mathbf{f}_{t,b})^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t \mathbf{D}_{t,b}\boldsymbol{\beta}])/\partial\boldsymbol{\beta} + \partial(\,\mathrm{E}[(\boldsymbol{\Upsilon}_t\mathbf{D}_{t,b}\boldsymbol{\beta})^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t \mathbf{f}_{t,b}])/\partial\boldsymbol{\beta}\Big)\end{aligned} \tag{115}$$

where

$$\boldsymbol{\Upsilon}_t = (\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I}_m) \tag{116}$$

Since $\boldsymbol{\beta}$ is to the far left or right in each term, the derivative is simple using the derivative terms in table 3.1. $\partial\Psi/\partial\boldsymbol{\beta}$ becomes:

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\upsilon} = -\frac{1}{2}\sum_{t=1}^{T}\Big(&- 2\,\mathrm{E}[\boldsymbol{X}_t^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t \mathbf{D}_{t,b}] + 2(\beta^\top \mathbf{D}_{t,b}^\top \boldsymbol{\Upsilon}_t^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t \mathbf{D}_{t,b}) \\ &+ 2\,\mathrm{E}[\mathbf{u}_t^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t \mathbf{D}_{t,b}] + 2\,\mathrm{E}[(\boldsymbol{\Upsilon}_t\mathbf{f}_{t,b})^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t \mathbf{D}_{t,b}]\Big)\end{aligned} \tag{117}$$

Note that $\boldsymbol{X}$ appears in $\boldsymbol{\Upsilon}_t$ but not in other terms. We need to keep track of where $\boldsymbol{X}$ appears so the we keep the expectation brackets around any terms involving $\boldsymbol{X}$.

$$\partial\Psi/\partial\boldsymbol{\beta} = \sum_{t=1}^{T}\Big(\,\mathrm{E}[\boldsymbol{X}_t^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t]\mathbf{D}_{t,b} - \mathbf{u}_t^\top \mathbb{Q}_t\,\mathrm{E}[\boldsymbol{\Upsilon}_t]\mathbf{D}_{t,b} - \boldsymbol{\beta}^\top \mathbf{D}_{t,b}^\top\,\mathrm{E}[\boldsymbol{\Upsilon}_t^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t]\mathbf{D}_{t,b} - \mathbf{f}_{t,b}^\top\,\mathrm{E}[\boldsymbol{\Upsilon}_t^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t]\mathbf{D}_{t,b}\Big) \tag{118}$$

Set the left side to zero and transpose the whole equation.

$$\mathbf{0} = \sum_{t=1}^{T}\Big(\mathbf{D}_{t,b}^\top\,\mathrm{E}[\boldsymbol{\Upsilon}_t^\top \mathbb{Q}_t \boldsymbol{X}_t] - \mathbf{D}_{t,b}^\top\,\mathrm{E}[\boldsymbol{\Upsilon}_t]^\top \mathbb{Q}_t \mathbf{u}_t - \mathbf{D}_{t,b}^\top\,\mathrm{E}[\boldsymbol{\Upsilon}_t^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t]\mathbf{f}_{t,b} - \mathbf{D}_{t,b}^\top\,\mathrm{E}[\boldsymbol{\Upsilon}_t^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t]\mathbf{D}_{t,b}\boldsymbol{\beta}\Big) \tag{119}$$

Thus,

$$(\sum_{t=1}^{T}\mathbf{D}_{t,b}^\top\,\mathrm{E}[\boldsymbol{\Upsilon}_t^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t]\mathbf{D}_{t,b})\boldsymbol{\beta} = \sum_{t=1}^{T}\mathbf{D}_{t,b}^\top(\,\mathrm{E}[\boldsymbol{\Upsilon}_t^\top \mathbb{Q}_t \boldsymbol{X}_t] - \,\mathrm{E}[\boldsymbol{\Upsilon}_t]^\top \mathbb{Q}_t \mathbf{u}_t - \,\mathrm{E}[\boldsymbol{\Upsilon}_t^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t]\mathbf{f}_{t,b}) \tag{120}$$

Now we need to deal with the expectations.

$$\begin{aligned}
\mathrm{E}[\boldsymbol{\Upsilon}_t^\top \mathbb{Q}_t \boldsymbol{\Upsilon}_t] &= \mathrm{E}[(\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I}_m)^\top \mathbb{Q}_t (\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I}_m)] \\
&= \mathrm{E}[(\boldsymbol{X}_{t-1} \otimes \mathbf{I}_m) \mathbb{Q}_t (\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I}_m)] \\
&= \mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top \otimes \mathbb{Q}_t] \\
&= \mathrm{E}[\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top] \otimes \mathbb{Q}_t \\
&= \widetilde{\mathbf{P}}_{t-1} \otimes \mathbb{Q}_t
\end{aligned} \tag{121}$$

$$\begin{aligned}
\mathrm{E}[\boldsymbol{\Upsilon}_t^\top \mathbb{Q}_t \boldsymbol{X}_t] &= \mathrm{E}[(\boldsymbol{X}_{t-1}^\top \otimes \mathbf{I}_m)^\top \mathbb{Q}_t \boldsymbol{X}_t] \\
&= \mathrm{E}[(\boldsymbol{X}_{t-1} \otimes \mathbf{I}_m) \mathbb{Q}_t \boldsymbol{X}_t] \\
&= \mathrm{E}[(\boldsymbol{X}_{t-1} \otimes \mathbb{Q}_t) \boldsymbol{X}_t] \\
&= \mathrm{E}[\,\mathrm{vec}(\mathbb{Q}_t \boldsymbol{X}_t \boldsymbol{X}_{t-1}^\top)] \\
&= \mathrm{vec}(\mathbb{Q}_t \widetilde{\mathbf{P}}_{t,t-1})
\end{aligned} \tag{122}$$

$$\begin{aligned}
\mathrm{E}[\boldsymbol{\Upsilon}_t]^\top \mathbb{Q}_t \mathbf{u}_t &= (\,\mathrm{E}[\boldsymbol{X}_{t-1}] \otimes \mathbf{I}_m) \mathbb{Q}_t \mathbf{u}_t \\
&= (\widetilde{\mathbf{x}}_{t-1} \otimes \mathbb{Q}_t) \mathbf{u}_t \\
&= \mathrm{vec}(\mathbb{Q}_t \mathbf{u}_t \widetilde{\mathbf{x}}_{t-1}^\top)
\end{aligned} \tag{123}$$

Thus,

$$\Big(\sum_{t=1}^{T} \mathbf{D}_{t,b}^\top (\widetilde{\mathbf{P}}_{t-1} \otimes \mathbb{Q}_t) \mathbf{D}_{t,b}\Big) \boldsymbol{\beta} = \sum_{t=1}^{T} \mathbf{D}_{t,b}^\top \big(\,\mathrm{vec}(\mathbb{Q}_t \widetilde{\mathbf{P}}_{t,t-1}) - (\widetilde{\mathbf{P}}_{t-1} \otimes \mathbb{Q}_t) \mathbf{f}_{t,b} - \,\mathrm{vec}(\mathbb{Q}_t \mathbf{u}_t \widetilde{\mathbf{x}}_{t-1}^\top)\big) \tag{124}$$

Then $\boldsymbol{\beta}$ for the $j+1$ iteration of the EM algorithm is then:

$$\boldsymbol{\beta} = \left(\sum_{t=1}^{T} \mathbf{D}_{t,b}^\top (\widetilde{\mathbf{P}}_{t-1} \otimes \mathbb{Q}_t) \mathbf{D}_{t,b}\right)^{-1} \times \sum_{t=1}^{T} \mathbf{D}_{t,b}^\top \big(\,\mathrm{vec}(\mathbb{Q}_t \widetilde{\mathbf{P}}_{t,t-1}) - (\widetilde{\mathbf{P}}_{t-1} \otimes \mathbb{Q}_t) \mathbf{f}_{t,b} - \,\mathrm{vec}(\mathbb{Q}_t \mathbf{u}_t \widetilde{\mathbf{x}}_{t-1}^\top)\big) \tag{125}$$

This requires that $\mathbf{D}_{t,b}^\top (\widetilde{\mathbf{P}}_{t-1} \otimes \mathbb{Q}_t) \mathbf{D}_{t,b}$ is invertible, and as usual we will run into trouble if $\Phi_t \mathbf{Q}_t \Phi_t^\top$ has zeros on the diagonal. See section 7.

5.7 The general Z update equation

The derivation of the update equation for $\boldsymbol{\zeta}$ with fixed and shared values is analogous to the derivation for $\boldsymbol{\beta}$. The update equation for $\boldsymbol{\zeta}$ is

$$\boldsymbol{\zeta}_{j+1} = \left(\sum_{t=1}^{T} \mathbf{D}_{t,z}^\top (\widetilde{\mathbf{P}}_t \otimes \mathbb{R}_t) \mathbf{D}_{t,z}\right)^{-1} \times \sum_{t=1}^{T} \mathbf{D}_{t,z}^\top \big(\,\mathrm{vec}(\mathbb{R}_t \widetilde{\mathbf{y}\mathbf{x}}_t) - (\widetilde{\mathbf{P}}_t \otimes \mathbb{R}_t) \mathbf{f}_{t,z} - \,\mathrm{vec}(\mathbb{R}_t \mathbf{a}_t \widetilde{\mathbf{x}}_t^\top)\big) \tag{126}$$

This requires that $\mathbf{D}_{t,z}^\top (\widetilde{\mathbf{P}}_t \otimes \mathbb{R}_t) \mathbf{D}_{t,z}$ is invertible. If $\Xi_t \mathbf{R}_t \Xi_t^\top$ has zeros on the diagonal, this will not be the case. See section 7.

5.8 The general Q update equation

A general analytical solution for $\mathbf{Q}$ is problematic because the inverse of $\mathbf{Q}_t$ appears in the likelihood and $\mathbf{Q}_t^{-1}$ cannot always be rewritten as a function of $\mathrm{vec}(\mathbf{Q}_t)$. 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 5.8.5). Indeed, all other cases (diagonal, block-diagonal, unconstrained, equal variance-covariance) except one (a replicated block-diagonal) are special cases of the blocked matrix with optional independent fixed blocks.

Unlike the other parameters, I need to put constraints on $\mathbf{f}$ and $\mathbf{D}$. I constrain $\mathbf{D}$ to be a design matrix. It has only 1s and 0s, and the rows sums are either 1 or 0. Thus terms like $q_1 + q_2$ are not allowed. A non-zero value in $\mathbf{f}$ is only allowed if the corresponding row in $\mathbf{D}$ is all zero. Thus elements like $f_1 + q_1$ are not allowed in $\mathbf{Q}$. These constraints, especially the constraint that $\mathbf{D}$ only has 0s and 1s, might be loosened, but with the addition of $\mathbf{G}_t$, we still have a very wide class of $\mathbf{Q}$ matrices.

The general update equation for $\mathbf{Q}$ with these constraints is

$$\begin{aligned}
\mathbf{q}_{j+1} &= \big(\sum_{t=1}^{T}(\mathbf{D}_{t,q}^\top \mathbf{D}_{t,q})\big)^{-1} \sum_{t=1}^{T} \mathbf{D}_{t,q}^\top \operatorname{vec}(\mathbf{S}_t) \\
\text{where } \mathbf{S}_t &= \Phi_t\big(\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}_t^\top - \mathbf{B}_t\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}_t^\top - \mathbf{u}_t\widetilde{\mathbf{x}}_t^\top + \\
&\qquad \mathbf{B}_t\widetilde{\mathbf{P}}_{t-1}\mathbf{B}_t^\top + \mathbf{B}_t\widetilde{\mathbf{x}}_{t-1}\mathbf{u}_t^\top + \mathbf{u}_t\widetilde{\mathbf{x}}_{t-1}^\top\mathbf{B}_t^\top + \mathbf{u}_t\mathbf{u}_t^\top\big)\Phi_t^\top \\
\operatorname{vec}(\mathbf{Q}_t)_{j+1} &= \mathbf{f}_{t,q} + \mathbf{D}_{t,q}\mathbf{q}_{j+1} \\
&\text{where} \\
\Phi_t &= (\mathbf{G}_t^\top \mathbf{G}_t)^{-1}\mathbf{G}_t^\top
\end{aligned} \tag{127}$$

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

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

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

Let $\mathbf{Q}$ be a non-time varying diagonal matrix with fixed and shared values such that it takes a form like so:

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

Here, f's are fixed values (constants) and q's are free parameters elements. The f and q do not occur together; i.e., there are no terms like $f_1 + q_1$.

The vec of $\mathbf{Q}^{-1}$ can be written then as $\operatorname{vec}(\mathbf{Q}^{-1}) = \mathbf{f}_q^* + \mathbf{D}_q\boldsymbol{q}^*$, 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.

$$\begin{aligned}
\partial\Psi/\partial\boldsymbol{q}^* = -\frac{1}{2}\sum_{t=1}^{T}\partial\Big(& \operatorname{E}[\boldsymbol{X}_t^\top\Phi_t^\top\mathbf{Q}^{-1}\Phi_t\boldsymbol{X}_t] - \operatorname{E}[\boldsymbol{X}_t^\top\Phi_t^\top\mathbf{Q}^{-1}\Phi_t\mathbf{B}_t\boldsymbol{X}_{t-1}] \\
&- \operatorname{E}[(\mathbf{B}_t\boldsymbol{X}_{t-1})^\top\Phi_t^\top\mathbf{Q}^{-1}\Phi_t\boldsymbol{X}_t] - \operatorname{E}[\boldsymbol{X}_t^\top\Phi_t^\top\mathbf{Q}^{-1}\Phi_t\mathbf{u}_t] \\
&- \operatorname{E}[\mathbf{u}_t^\top\Phi_t^\top\mathbf{Q}^{-1}\Phi_t\boldsymbol{X}_t] + \operatorname{E}[(\mathbf{B}_t\boldsymbol{X}_{t-1})^\top\Phi_t^\top\mathbf{Q}^{-1}\Phi_t\mathbf{B}_t\boldsymbol{X}_{t-1}] \\
&+ \operatorname{E}[(\mathbf{B}_t\boldsymbol{X}_{t-1})^\top\Phi_t^\top\mathbf{Q}^{-1}\Phi_t\mathbf{u}_t] + \operatorname{E}[\mathbf{u}_t^\top\Phi_t^\top\mathbf{Q}^{-1}\Phi_t\mathbf{B}_t\boldsymbol{X}_{t-1}] + \mathbf{u}_t^\top\Phi_t^\top\mathbf{Q}^{-1}\Phi_t\mathbf{u}_t\Big)/\partial\boldsymbol{q}^* \\
&- \partial\big(\frac{T}{2}\log|\mathbf{Q}|\big)/\partial\boldsymbol{q}^*
\end{aligned} \tag{128}$$

Use vec operation Equation 82 to pull $\mathbf{Q}^{-1}$ out from the middle[13], using

$$\mathbf{a}^\top \Phi^\top \mathbf{Q}^{-1} \Phi \mathbf{b} = (\mathbf{b}^\top \Phi^\top \otimes \mathbf{a}^\top \Phi^\top)\,\mathrm{vec}(\mathbf{Q}^{-1}) = (\mathbf{b}^\top \otimes \mathbf{a}^\top)(\Phi^\top \otimes \Phi^\top)\,\mathrm{vec}(\mathbf{Q}^{-1})$$

. Then replace the expectations with the Kalman smoother output,

$$\begin{aligned}\partial\Psi/\partial \boldsymbol{q}^* = -\frac{1}{2}\sum_{t=1}^{T} \partial\Big(&\mathrm{E}[\boldsymbol{X}_t^\top \otimes \boldsymbol{X}_t^\top] - \mathrm{E}[\boldsymbol{X}_t^\top \otimes (\mathbf{B}_t \boldsymbol{X}_{t-1})^\top] - \mathrm{E}[(\mathbf{B}_t \boldsymbol{X}_{t-1})^\top \otimes \boldsymbol{X}_t^\top] \\ &- \mathrm{E}[\boldsymbol{X}_t^\top \otimes \mathbf{u}_t^\top] - \mathrm{E}[\mathbf{u}_t^\top \otimes \boldsymbol{X}_t^\top] + \mathrm{E}[(\mathbf{B}_t \boldsymbol{X}_{t-1})^\top \otimes (\mathbf{B}_t \boldsymbol{X}_{t-1})^\top] \\ &+ \mathrm{E}[(\mathbf{B}_t \boldsymbol{X}_{t-1})^\top \otimes \mathbf{u}_t^\top] + \mathrm{E}[\mathbf{u}_t^\top \otimes (\mathbf{B}\boldsymbol{X}_{t-1})^\top] + (\mathbf{u}_t^\top \otimes \mathbf{u}_t^\top)\Big)(\Phi_t \otimes \Phi_t)^\top\,\mathrm{vec}(\mathbf{Q}^{-1})/\partial \boldsymbol{q}^* \\ &- \partial\Big(\frac{T}{2}\log|\mathbf{Q}|\Big)/\partial \boldsymbol{q}^*\end{aligned} \tag{129}$$

This can be further reduced using

$$(\mathbf{b}^\top \otimes \mathbf{a}^\top)(\Phi^\top \otimes \Phi^\top) = (\,\mathrm{vec}(\mathbf{a}\mathbf{b}^\top))^\top (\Phi \otimes \Phi)^\top = \,\mathrm{vec}(\Phi \mathbf{a}\mathbf{b}^\top \Phi^\top)^\top$$

With this reduction and replacing $\log|\mathbf{Q}|$ with $-\log|\mathbf{Q}^{-1}|$, we get

$$\begin{aligned}&\partial\Psi/\partial \boldsymbol{q}^* = -\frac{1}{2}\sum_{t=1}^{T} \mathrm{vec}(\mathbf{S}_t)^\top \partial\big(\,\mathrm{vec}(\mathbf{Q}^{-1})\big)/\partial \boldsymbol{q}^* + \partial\big(\frac{T}{2}\log|\mathbf{Q}^{-1}|\big)/\partial \boldsymbol{q}^* \\ &\text{where} \\ &\mathbf{S}_t = \Phi_t\big(\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}_t^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t \mathbf{u}_t^\top - \mathbf{u}_t \widetilde{\mathbf{x}}_t^\top + \\ &\quad \mathbf{B}_t \widetilde{\mathbf{P}}_{t-1}\mathbf{B}_t^\top + \mathbf{B}_t \widetilde{\mathbf{x}}_{t-1}\mathbf{u}_t^\top + \mathbf{u}_t \widetilde{\mathbf{x}}_{t-1}^\top \mathbf{B}_t^\top + \mathbf{u}_t \mathbf{u}_t^\top\big)\Phi_t^\top\end{aligned} \tag{130}$$

The determinant of a diagonal matrix is the product of its diagonal elements. Thus,

$$\begin{aligned}\partial\Psi/\partial \boldsymbol{q}^* = -\Big(&\frac{1}{2}\sum_{t=1}^{T} \mathrm{vec}(\mathbf{S}_t)^\top (\mathbf{f}^* + \mathbf{D}_q \boldsymbol{q}^*) \\ &- \frac{1}{2}\sum_{t=1}^{T} (\log(f_1^*) + \log(f_2^*)...k\log(q_1^*) + l\log(q_2^*)...)\Big)/\partial \boldsymbol{q}^*\end{aligned} \tag{131}$$

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 and transposing the whole equation we get,

$$\begin{aligned}\partial\Psi/\partial \boldsymbol{q}^* &== \frac{1}{2}\sum_{t=1}^{T} \mathbf{D}_q^\top\,\mathrm{vec}(\mathbf{S}_t) - \frac{1}{2}\sum_{t=1}^{T} (\log(f_1^*) + ...k\log(q_1^*) + l\log(q_2^*)...)/\partial \boldsymbol{q}^* \\ &= \frac{1}{2}\sum_{t=1}^{T} \mathbf{D}_q^\top\,\mathrm{vec}(\mathbf{S}_t) - \frac{1}{2}\sum_{t=1}^{T} \mathbf{D}_q^\top \mathbf{D}_q \boldsymbol{q}\end{aligned} \tag{132}$$

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

$$\begin{aligned}\mathbf{q}_{j+1} &= \big(\sum_{t=1}^{T} \mathbf{D}_q^\top \mathbf{D}_q\big)^{-1} \sum_{t=1}^{T} \mathbf{D}_q^\top\,\mathrm{vec}(\mathbf{S}_t) \\ \mathrm{vec}(\mathbf{Q})_{j+1} &= \mathbf{f} + \mathbf{D}_q \boldsymbol{q}_{j+1}\end{aligned} \tag{133}$$

Since in this example, $\mathbf{D}_q$ is time-constant, this reduces to

$$\boldsymbol{q}_{j+1} = \frac{1}{T}(\mathbf{D}_q^\top \mathbf{D}_q)^{-1} \mathbf{D}_q^\top \sum_{t=1}^{T} \mathrm{vec}(\mathbf{S}_t)$$

$\mathbf{S}_t$ is defined in equation 129.

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

5.8.2 Special case: Q with one variance and one covariance

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

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

$$\partial\Psi/\partial\boldsymbol{q}^{*} = \partial\Big(-\frac{1}{2}\sum_{t=1}^{T}\big(\operatorname{vec}(\mathbf{S}_t)^{\top}\big)\operatorname{vec}(\mathbf{Q}^{-1}) + \frac{T}{2}\log|\mathbf{Q}^{-1}|\Big)/\partial\boldsymbol{q}^{*} \tag{134}$$

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:

$$\partial\Psi/\partial\boldsymbol{q}^{*} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{D}^{\top}\operatorname{vec}(\mathbf{S}_t) - \frac{1}{2}\sum_{t=1}^{T}(\mathbf{D}^{\top}\mathbf{D})\boldsymbol{q} \tag{135}$$

where here $\mathbf{D}^{\top}\mathbf{D}$ is a 2×2 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 $\mathbf{D}$; $s=2$ 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 133 in which $\mathbf{f}_q = 0$ since there are no fixed values.

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

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

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

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

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

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

This is ok:

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

This is not ok:

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

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

5.8.4 Special case: a symmetric blocked matrix

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

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

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

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

The following are block-symmetric:

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

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

The following are NOT legal block-symmetric matrices:

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

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

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 ($\mathbb{E}$).

5.8.5 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).

5.9 The general R update equation

The $\mathbf{R}$ update equation for blocked symmetric matrices with optional independent fixed blocks is completely analogous to the $\mathbf{Q}$ equation. Thus if $\mathbf{R}$ 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. Elements like $f_i + r_j$ and $r_i + r_j$ are not allowed in $\mathbf{R}$. Only elements of the form f_i and r_i are allowed. If an element has a fixed component, it must be completely fixed. Each element in $\mathbf{R}$ can have only one of the elements in $\mathbf{r}$, but multiple elements in $\mathbf{R}$ can have the same $\mathbf{r}$ element.

The update equation is

$$\begin{aligned} \mathbf{r}_{j+1} &= \left(\sum_{t=1}^{T} \mathbf{D}_{t,r}^\top \mathbf{D}_{t,r} \right)^{-1} \sum_{t=1}^{T} \mathbf{D}_{t,r}^\top \operatorname{vec}\left(\mathbf{T}_{t,j+1} \right) \\ \operatorname{vec}(\mathbf{R}_t)_{j+1} &= \mathbf{f}_{t,r} + \mathbf{D}_{t,r} \mathbf{r}_{j+1} \end{aligned} \tag{136}$$

The $\mathbf{T}_{t,j+1}$ used at time step t in equation 136 is the term that appears in the summation in the unconstrained update equation with no missing values (equation 54):

$$\mathbf{T}_{t,j+1} = \Xi_t \left(\widetilde{\mathbf{O}}_t - \widetilde{\mathbf{y}\mathbf{x}}_t \mathbf{Z}_t^\top - \mathbf{Z}_t \widetilde{\mathbf{y}\mathbf{x}}_t^\top - \widetilde{\mathbf{y}}_t \mathbf{a}_t^\top - \mathbf{a}_t \widetilde{\mathbf{y}}_t^\top + \mathbf{Z}_t \widetilde{\mathbf{P}}_t \mathbf{Z}_t^\top + \mathbf{Z}_t \widetilde{\mathbf{x}}_t \mathbf{a}_t^\top + \mathbf{a}_t \widetilde{\mathbf{x}}_t^\top \mathbf{Z}_t^\top + \mathbf{a}_t \mathbf{a}_t^\top \right) \Xi_t^\top \tag{137}$$

where $\Xi_t = (\mathbf{H}_t^\top \mathbf{H}_t)^{-1} \mathbf{H}_t^\top$.

6 Computing the expectations in the update equations

For the update equations, we need to compute the expectations of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ and their products conditioned on 1) the observed data $\boldsymbol{Y}(1) = \boldsymbol{y}(1)$ and 2) the parameters at time t, Θ_j. This section shows how to compute these expectations. Throughout the section, I will normally leave off the conditional $\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j$ when specifying an expectation. Thus any E[] appearing without its conditional is conditioned on $\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j$.

However if there are additional or different conditions those will be shown. Also all expectations are over the joint distribution of XY unless explicitly specified otherwise.

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

$$\boldsymbol{y}_t = \begin{bmatrix} y_1 \\ NA \\ y_3 \\ y_4 \\ NA \\ y_6 \end{bmatrix} \tag{138}$$

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

Let $\boldsymbol{\Omega}_t^{(1)}$ 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,

$$\begin{aligned} \boldsymbol{Y}_t(1) &= \boldsymbol{\Omega}_t^{(1)} \boldsymbol{Y}_t, \quad \boldsymbol{\Omega}_t^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \\ \boldsymbol{Y}_t(2) &= \boldsymbol{\Omega}_t^{(2)} \boldsymbol{Y}_t, \quad \boldsymbol{\Omega}_t^{(2)} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{aligned} \tag{139}$$

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,

$$\begin{aligned} \mathbf{I}_t^{(1)} &= (\boldsymbol{\Omega}_t^{(1)})^\top \boldsymbol{\Omega}_t^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \\ \mathbf{I}_t^{(2)} &= (\boldsymbol{\Omega}_t^{(2)})^\top \boldsymbol{\Omega}_t^{(2)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \end{aligned} \tag{140}$$

6.1 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. The Kalman filter provides the expectations involving only $\boldsymbol{X}_t$ conditioned on all the data from time 1 to $t-1$ and from time 1 to t. For the EM algorithm, we only need the smoother output and the expected values conditioned on the data from time 1 to T and these are denoted with special symbol of a tilde over a variable.

To present the algorithm for the Kalman smoother and filter, the expectations conditioned on time 1 to t are needed. The notation for this general case will be $\boldsymbol{x}_1^t$ to denote $\mathrm{E}[\boldsymbol{X}_t | \boldsymbol{Y}(1)_1^t = \boldsymbol{y}(1)_1^t, \Theta]$ where $\boldsymbol{y}(1)_1^t$ means the observed data (the (1) part) from time 1 to t (the superscript). This is fairly common notation for the conditional expectations in a Kalman filter and smoother and it is important to note that the superscript is not a power notation but the upper time extent. The the expectations used in the previous sections on the EM algorithm are the following:

$$\widetilde{\mathbf{x}}_t \equiv \boldsymbol{x}_t^T = \mathrm{E}[\boldsymbol{X}_t|\boldsymbol{Y}(1)_1^T = \boldsymbol{y}(1)_1^T, \Theta] \tag{141a}$$

$$\widetilde{\boldsymbol{V}}_t \equiv \boldsymbol{V}_t^T = \mathrm{var}[\boldsymbol{X}_t|\boldsymbol{Y}(1)_1^T = \boldsymbol{y}(1)_1^T, \Theta] \tag{141b}$$

$$\widetilde{\boldsymbol{V}}_{t,t-1} \equiv \boldsymbol{V}_{t,t-1}^T = \mathrm{cov}[\boldsymbol{X}_t, \boldsymbol{X}_{t-1}|\boldsymbol{Y}(1)_1^T = \boldsymbol{y}(1)_1^T, \Theta] \tag{141c}$$

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

$$\widetilde{\mathbf{P}}_t \equiv \mathbf{P}_t^T = \mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)_1^T = \boldsymbol{y}(1)_1^T, \Theta] = \widetilde{\boldsymbol{V}}_t + \widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^\top \tag{141d}$$

$$\widetilde{\mathbf{P}}_{t,t-1} \equiv \mathbf{P}_{t,t-1}^T = \mathrm{E}[\boldsymbol{X}_t\boldsymbol{X}_{t-1}^\top|\boldsymbol{Y}(1)_1^T = \boldsymbol{y}(1)_1^T, \Theta] = \widetilde{\boldsymbol{V}}_{t,t-1} + \widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_{t-1}^\top \tag{141e}$$

The $\widetilde{\mathbf{P}}_t$ and $\widetilde{\mathbf{P}}_{t,t-1}$ equations arise from the computational formula for variance (equation 12). When comparing the Kalman filter and smoother algorithms here to Shumway and Stoffer, keep in mind the difference in notation: P_t^n in Shumway and Stoffer is $\boldsymbol{V}_t^T$ here not $\mathbf{P}_t^T$.

In the presentation of the EM algorithm, $\boldsymbol{Y}(1)_1^T = \boldsymbol{y}(1)_1^T, \Theta$ was dropped from the expectations to remove clutter; thus $E[\ldots]$ always denoted the conditional expectation $\mathrm{E}[\ldots|\boldsymbol{Y}(1)_1^T = \boldsymbol{y}(1)_1^T, \Theta]$. To present the smoother algorithm, I present the other conditional expectations.

$$\boldsymbol{x}_t^{t-1} = \mathrm{E}[\boldsymbol{X}_t|\boldsymbol{Y}(1)_1^{t-1} = \boldsymbol{y}(1)_1^{t-1}, \Theta] \tag{142a}$$

$$\boldsymbol{x}_t^t = \mathrm{E}[\boldsymbol{X}_t|\boldsymbol{Y}(1)_1^t = \boldsymbol{y}(1)_1^t, \Theta] \tag{142b}$$

$$\boldsymbol{V}_t^{t-1} = \mathrm{var}[\boldsymbol{X}_t|\boldsymbol{Y}(1)_1^{t-1} = \boldsymbol{y}(1)_1^{t-1}, \Theta] \tag{142c}$$

$$\boldsymbol{V}_t^t = \mathrm{var}[\boldsymbol{X}_t|\boldsymbol{Y}(1)_1^t = \boldsymbol{y}(1)_1^t, \Theta] \tag{142d}$$

$$\tag{142e}$$

The first part of the Kalman smoother algorithm is the Kalman filter which gives the expectation at time t conditioned on the data up to time t. The following the filter as shown in (Shumway and Stoffer, 2006, section 6.2, p. 331), although the notation is a little different. The recursion starts at time $t = 1$ and repeats until $t = T$.

$$\boldsymbol{x}_t^{t-1} = \mathbf{B}_t\boldsymbol{x}_{t-1}^{t-1} + \mathbf{u}_t \tag{143a}$$

$$\boldsymbol{V}_t^{t-1} = \mathbf{B}_t\boldsymbol{V}_{t-1}^{t-1}\mathbf{B}_t^\top + \mathbf{G}_t\mathbf{Q}_t\mathbf{G}_t^\top \tag{143b}$$

$$\boldsymbol{x}_t^t = \boldsymbol{x}_t^{t-1} + \mathbf{K}_t(\boldsymbol{y}_t - \mathbf{Z}_t\boldsymbol{x}_t^{t-1} - \mathbf{a}_t) \tag{143c}$$

$$\boldsymbol{V}_t^t = (\mathbf{I}_m - \mathbf{K}_t\mathbf{Z}_t)\boldsymbol{V}_t^{t-1} \tag{143d}$$

$$\mathbf{K}_t = \boldsymbol{V}_t^{t-1}\mathbf{Z}_t^\top(\mathbf{Z}_t\boldsymbol{V}_t^{t-1}\mathbf{Z}_t^\top + \mathbf{H}_t\mathbf{R}_t\mathbf{H}_t^\top)^{-1} \tag{143e}$$

If the initial value is defined at $t = 0$, then the filter starts at $t = 1$ with the first two equations with $\boldsymbol{x}_0^0 \equiv \boldsymbol{\xi}$ and $\boldsymbol{V}_0 \equiv \boldsymbol{\Lambda}$. If the initial value is defined at $t = 1$, then the filter starts at $t = 1$ with the third and fourth equations with $\boldsymbol{x}_1^0 \equiv \boldsymbol{\xi}$ and $\boldsymbol{V}_1^0 \equiv \boldsymbol{\Lambda}$.

The Kalman smoother and lag-1 covariance smoother compute the expectations conditioned on all the data, 1 to T:

$$\boldsymbol{x}_{t-1}^T = \boldsymbol{x}_{t-1}^{t-1} + \mathbf{J}_{t-1}(\boldsymbol{x}_t^T - \boldsymbol{x}_t^{t-1}) \tag{144a}$$

$$\boldsymbol{V}_{t-1}^T = \boldsymbol{V}_{t-1}^{t-1} + \mathbf{J}_{t-1}(\boldsymbol{V}_t^T - \boldsymbol{V}_t^{t-1})\mathbf{J}_{t-1}^\top \tag{144b}$$

$$\mathbf{J}_{t-1} = \boldsymbol{V}_{t-1}^{t-1}\mathbf{B}_t^\top(\boldsymbol{V}_t^{t-1})^{-1} \tag{144c}$$

$$\tag{144d}$$

$$\boldsymbol{V}_{T,T-1}^T = (\mathbf{I} - \mathbf{K}_T\mathbf{Z}_T)\mathbf{B}_T\boldsymbol{V}_{T-1}^{T-1} \tag{144e}$$

$$\boldsymbol{V}_{t-1,t-2}^T = \boldsymbol{V}_{t-1}^{t-1}\mathbf{J}_{t-2}^\top + \mathbf{J}_{t-1}((\boldsymbol{V}_{t,t-1}^T - \mathbf{B}_t\boldsymbol{V}_{t-1}^{t-1}))\mathbf{J}_{t-2}^\top \tag{144f}$$

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, section 6.4, eqn 6.78, p. 348). In this case, the usual filter and smoother equations are used with the following modifications to the parameters and data

used in the equations. If the i-th element of $\boldsymbol{y}_t$ is missing, zero out the i-th rows in $\boldsymbol{y}_t$, $\mathbf{a}$ and $\mathbf{Z}$. Thus if the 2nd and 5th elements of $\boldsymbol{y}_t$ are missing,

$$\boldsymbol{y}_t^* = \begin{bmatrix} y_1 \\ 0 \\ y_3 \\ y_4 \\ 0 \\ y_6 \end{bmatrix}, \quad \mathbf{a}_t^* = \begin{bmatrix} a_1 \\ 0 \\ a_3 \\ a_4 \\ 0 \\ a_6 \end{bmatrix}, \quad \mathbf{Z}_t^* = \begin{bmatrix} z_{1,1} & z_{1,2} & \dots \\ 0 & 0 & \dots \\ z_{3,1} & z_{3,2} & \dots \\ z_{4,1} & z_{4,2} & \dots \\ 0 & 0 & \dots \\ z_{6,1} & z_{6,2} & \dots \end{bmatrix} \tag{145}$$

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

$$\mathbf{R}_t = \mathbf{H}_t \mathbf{R} \mathbf{H}_t^\top = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} & r_{1,5} & r_{1,6} \\ r_{2,1} & r_{2,2} & r_{2,3} & r_{2,4} & r_{2,5} & r_{2,6} \\ r_{3,1} & r_{3,2} & r_{3,3} & r_{3,4} & r_{3,5} & r_{3,6} \\ r_{4,1} & r_{4,2} & r_{4,3} & r_{4,4} & r_{4,5} & r_{4,6} \\ r_{5,1} & r_{5,2} & r_{5,3} & r_{5,4} & r_{5,5} & r_{5,6} \\ r_{6,1} & r_{6,2} & r_{6,3} & r_{6,4} & r_{6,5} & r_{6,6} \end{bmatrix} \tag{146}$$

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

$$\mathbf{R}_t^* = \begin{bmatrix} r_{1,1} & 0 & r_{1,3} & r_{1,4} & 0 & r_{1,6} \\ 0 & r_{2,2} & 0 & 0 & r_{2,5} & 0 \\ r_{3,1} & 0 & r_{3,3} & r_{3,4} & 0 & r_{3,6} \\ r_{4,1} & 0 & r_{4,3} & r_{4,4} & 0 & r_{4,6} \\ 0 & r_{5,2} & 0 & 0 & r_{5,5} & 0 \\ r_{6,1} & 0 & r_{6,3} & r_{6,4} & 0 & r_{6,6} \end{bmatrix} \tag{147}$$

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

$$\begin{aligned} \boldsymbol{y}_t^* &= \mathbf{I}_t^{(1)} \boldsymbol{y}_t \\ \mathbf{a}_t^* &= \mathbf{I}_t^{(1)} \mathbf{a}_t \\ \mathbf{Z}_t^* &= \mathbf{I}_t^{(1)} \mathbf{Z}_t \\ \mathbf{R}_t^* &= \mathbf{I}_t^{(1)} \mathbf{R}_t \mathbf{I}_t^{(1)} + \mathbf{I}_t^{(2)} \mathbf{R}_t \mathbf{I}_t^{(2)} \end{aligned} \tag{148}$$

$\mathbf{a}_t^*$, $\mathbf{Z}_t^*$ 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.

6.2 Expectations involving $\boldsymbol{Y}_t$

First, replace the missing values in $\boldsymbol{y}_t$ with zeros[14] and then the expectations are given by the following equations. The derivations for these equations are given in the subsections to follow.

$$\widetilde{\mathbf{y}}_t = \mathrm{E}[\boldsymbol{Y}_t] = \boldsymbol{y}_t - \nabla_t(\boldsymbol{y}_t - \mathbf{Z}_t \widetilde{\mathbf{x}}_t - \mathbf{a}_t) \tag{149a}$$

$$\widetilde{\mathbf{O}}_t = \mathrm{E}[\boldsymbol{Y}_t \boldsymbol{Y}_t^\top] = \mathbf{I}_t^{(2)}(\nabla_t \mathbf{H}_t \mathbf{R}_t \mathbf{H}_t^\top + \nabla_t \mathbf{Z}_t \widetilde{\boldsymbol{V}}_t \mathbf{Z}_t^\top \nabla_t^\top)\mathbf{I}_t^{(2)} + \widetilde{\mathbf{y}}_t \widetilde{\mathbf{y}}_t^\top \tag{149b}$$

$$\widetilde{\mathbf{y}\mathbf{x}}_t = \mathrm{E}[\boldsymbol{Y}_t \boldsymbol{X}_t^\top] = \nabla_t \mathbf{Z}_t \widetilde{\boldsymbol{V}}_t + \widetilde{\mathbf{y}}_t \widetilde{\mathbf{x}}_t^\top \tag{149c}$$

$$\widetilde{\mathbf{y}\mathbf{x}}_{t,t-1} = \mathrm{E}[\boldsymbol{Y}_t \boldsymbol{X}_{t-1}^\top] = \nabla_t \mathbf{Z}_t \widetilde{\boldsymbol{V}}_{t,t-1} + \widetilde{\mathbf{y}}_t \widetilde{\mathbf{x}}_{t-1}^\top \tag{149d}$$

$$\text{where } \nabla_t = \mathbf{I} - \mathbf{H}_t \mathbf{R}_t \mathbf{H}_t^\top (\boldsymbol{\Omega}_t^{(1)})^\top (\boldsymbol{\Omega}_t^{(1)} \mathbf{H}_t \mathbf{R}_t \mathbf{H}_t^\top (\boldsymbol{\Omega}_t^{(1)})^\top)^{-1} \boldsymbol{\Omega}_t^{(1)} \tag{149e}$$

$$\text{and } \mathbf{I}_t^{(2)} = (\boldsymbol{\Omega}_t^{(2)})^\top \boldsymbol{\Omega}_t^{(2)} \tag{149f}$$

[14] 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 $\boldsymbol{y}_t$ 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 $\mathbf{R}_t$ is diagonal, then $\mathbf{R}_t(\boldsymbol{\Omega}_t^{(1)})^\top(\boldsymbol{\Omega}_t^{(1)}\mathbf{R}_t(\boldsymbol{\Omega}_t^{(1)})^\top)^{-1}\boldsymbol{\Omega}_t^{(1)} = \mathbf{I}_t^{(1)}$ and $\nabla_t = \mathbf{I}_t^{(2)}$. This will mean that in $\widetilde{\mathbf{y}}_t$ the $\boldsymbol{y}_t(2)$ are given by $\mathbf{Z}_t\widetilde{\mathbf{x}}_t + \mathbf{a}_t$, as expected when $\boldsymbol{y}_t(1)$ and $\boldsymbol{y}_t(2)$ are independent.

If there are zeros on the diagonal of $\mathbf{R}_t$ (section 7), the definition of ∇_t is changed slightly from that shown in equation 149. Let $\mho_t^{(r)}$ be the matrix that extracts the elements of $\boldsymbol{y}_t$ where $\boldsymbol{y}_t(i)$ is not missing AND $\mathbf{H}_t\mathbf{R}_t(i,i)\mathbf{H}_t^\top$ is not zero. Then

$$\nabla_t = \mathbf{I} - \mathbf{H}_t\mathbf{R}_t\mathbf{H}_t^\top(\mho_t^{(r)})^\top(\mho_t^{(r)}\mathbf{H}_t\mathbf{R}_t\mathbf{H}_t^\top(\mho_t^{(r)})^\top)^{-1}\mho_t^{(r)} \tag{150}$$

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

In the MARSS equation, the observation errors are denoted $\mathbf{H}_t\boldsymbol{v}_t$. $\boldsymbol{v}_t$ is a specific realization from a random variable $\boldsymbol{V}_t$ that is distributed multivariate normal with mean 0 and variance $\mathbf{R}_t$. $\boldsymbol{V}_t$ is not to be confused with $\widetilde{\boldsymbol{V}}_t$ in equation 141, which is unrelated[15] to $\boldsymbol{V}_t$. If there are no missing values, then we condition on $\boldsymbol{Y}_t = \boldsymbol{y}_t$ and

$$\mathrm{E}[\boldsymbol{Y}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1)] = \mathrm{E}[\boldsymbol{Y}_t|\boldsymbol{Y}_t = \boldsymbol{y}_t] = \boldsymbol{y}_t \tag{151}$$

If there are no observed values, then

$$\mathrm{E}[\boldsymbol{Y}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1)] = \mathrm{E}[\boldsymbol{Y}_t] = \mathrm{E}[\mathbf{Z}_t\boldsymbol{X}_t + \mathbf{a}_t + \boldsymbol{V}_t] = \mathbf{Z}_t\widetilde{\mathbf{x}}_t + \mathbf{a}_t \tag{152}$$

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):

$$\text{If, } \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \tag{153}$$

Then,

$$\begin{aligned} &(Y_1|Y_1 = y_1) = y_1, \quad \text{and} \\ &(Y_2|Y_1 = y_1) \sim \mathrm{MVN}(\bar{\mu}, \bar{\Sigma}), \quad \text{where} \\ &\qquad \bar{\mu} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_1) \\ &\qquad \bar{\Sigma} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned} \tag{154}$$

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

$$\begin{bmatrix} \boldsymbol{Y}_t(1)|\boldsymbol{X}_t = \boldsymbol{x}_t \\ \boldsymbol{Y}_t(2)|\boldsymbol{X}_t = \boldsymbol{x}_t \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} \boldsymbol{\Omega}_t^{(1)}(\mathbf{Z}_t\boldsymbol{x}_t + \mathbf{a}_t) \\ \boldsymbol{\Omega}_t^{(2)}(\mathbf{Z}_t\boldsymbol{x}_t + \mathbf{a}_t) \end{bmatrix}, \begin{bmatrix} (\mathbf{H}_t\mathbf{R}_t\mathbf{H}_t^\top)_{11} & (\mathbf{H}_t\mathbf{R}_t\mathbf{H}_t^\top)_{12} \\ (\mathbf{H}_t\mathbf{R}_t\mathbf{H}_t^\top)_{21} & (\mathbf{H}_t\mathbf{R}_t\mathbf{H}_t^\top)_{22} \end{bmatrix}\right) \tag{155}$$

Thus,

$$\begin{aligned} &(\boldsymbol{Y}_t(1)|\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1), \boldsymbol{X}_t = \boldsymbol{x}_t) = \boldsymbol{\Omega}_t^{(1)}\boldsymbol{y}_t \quad \text{and} \\ &(\boldsymbol{Y}_t(2)|\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1), \boldsymbol{X}_t = \boldsymbol{x}_t) \sim \mathrm{MVN}(\ddot{\mu}, \ddot{\Sigma}) \quad \text{where} \\ &\quad \ddot{\mu} = \boldsymbol{\Omega}_t^{(2)}(\mathbf{Z}_t\boldsymbol{x}_t + \mathbf{a}_t) + \ddot{\mathbf{R}}_{t,21}(\ddot{\mathbf{R}}_{t,11})^{-1}\boldsymbol{\Omega}_t^{(1)}(\boldsymbol{y}_t - \mathbf{Z}_t\boldsymbol{x}_t - \mathbf{a}_t) \\ &\quad \ddot{\Sigma} = \ddot{\mathbf{R}}_{t,22} - \ddot{\mathbf{R}}_{t,21}(\ddot{\mathbf{R}}_{t,11})^{-1}\ddot{\mathbf{R}}_{t,12} \\ &\quad \ddot{\mathbf{R}}_t = \mathbf{H}_t\mathbf{R}_t\mathbf{H}_t^\top \end{aligned} \tag{156}$$

Note that since we are conditioning on $\boldsymbol{X}_t = \boldsymbol{x}_t$, we can replace $\boldsymbol{Y}$ (all data from time 1 to T) by $\boldsymbol{Y}_t$ (data at time t) in the conditional:

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

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

From this and the distributions in equation 156, we can write down $\widetilde{\mathbf{y}}_t = \mathrm{E}[\boldsymbol{Y}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1), \Theta_j]$:

$$\begin{aligned}
\widetilde{\mathbf{y}}_t &= \mathrm{E}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}(1) = \boldsymbol{y}(1)] \\
&= \int_{\boldsymbol{x}_t}\int_{\boldsymbol{y}_t} \boldsymbol{y}_t f(\boldsymbol{y}_t|\boldsymbol{y}_t(1), \boldsymbol{x}_t) d\boldsymbol{y}_t f(\boldsymbol{x}_t) d\boldsymbol{x}_t \\
&= \mathrm{E}_X[\,\mathrm{E}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1), \boldsymbol{X}_t = \boldsymbol{x}_t]] \\
&= \mathrm{E}_X[\boldsymbol{y}_t - \nabla_t(\boldsymbol{y}_t - \mathbf{Z}_t\boldsymbol{X}_t - \mathbf{a}_t)] \\
&= \boldsymbol{y}_t - \nabla_t(\boldsymbol{y}_t - \mathbf{Z}_t\widetilde{\mathbf{x}}_t - \mathbf{a}_t) \\
\text{where } \nabla_t &= \mathbf{I} - \ddot{\mathbf{R}}_t(\boldsymbol{\Omega}_t^{(1)})^\top(\ddot{\mathbf{R}}_{t,11})^{-1}\boldsymbol{\Omega}_t^{(1)}
\end{aligned} \tag{157}$$

$(\boldsymbol{\Omega}_t^{(1)})^\top(\ddot{\mathbf{R}}_{t,11})^{-1}\boldsymbol{\Omega}_t^{(1)}$ 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 $\mathbf{I}$.

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

The following outlines a[16] derivation. If there are no missing values, then we condition on $\boldsymbol{Y}_t = \boldsymbol{y}_t$ and

$$\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}(1) = \boldsymbol{y}(1)] = \mathrm{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}_t = \boldsymbol{y}_t] = \boldsymbol{y}_t\boldsymbol{y}_t^\top. \tag{158}$$

If there are no observed values at time t, then

$$\begin{aligned}
&\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top] \\
&\quad= \mathrm{var}[\mathbf{Z}_t\boldsymbol{X}_t + \mathbf{a}_t + \mathbf{H}_t\boldsymbol{V}_t] + \mathrm{E}[\mathbf{Z}_t\boldsymbol{X}_t + \mathbf{a}_t + \mathbf{H}_t\boldsymbol{V}_t]\,\mathrm{E}[\mathbf{Z}_t\boldsymbol{X}_t + \mathbf{a}_t + \mathbf{H}_t\boldsymbol{V}_t]^\top \\
&\quad= \mathrm{var}[\boldsymbol{V}_t] + \mathrm{var}[\mathbf{Z}_t\boldsymbol{X}_t] + (\,\mathrm{E}[\mathbf{Z}_t\boldsymbol{X}_t + \mathbf{a}_t] + \mathrm{E}[\mathbf{H}_t\boldsymbol{V}_t])(\,\mathrm{E}[\mathbf{Z}_t\boldsymbol{X}_t + \mathbf{a}_t] + \mathrm{E}[\mathbf{H}_t\boldsymbol{V}_t])^\top \\
&\quad= \ddot{\mathbf{R}}_t + \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t\mathbf{Z}_t^\top + (\mathbf{Z}_t\widetilde{\mathbf{x}}_t + \mathbf{a}_t)(\mathbf{Z}_t\widetilde{\mathbf{x}}_t + \mathbf{a}_t)^\top
\end{aligned} \tag{159}$$

When only some of the $\boldsymbol{Y}_t$ are observed, we use again the conditional probability of a multivariate normal (equation 153). From this property, we know that

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

$$\begin{aligned}
\text{Thus } &\mathrm{var}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1), \boldsymbol{X}_t = \boldsymbol{x}_t] \\
&= (\boldsymbol{\Omega}_t^{(2)})^\top(\ddot{\mathbf{R}}_{t,22} - \ddot{\mathbf{R}}_{t,21}(\ddot{\mathbf{R}}_{t,11})^{-1}\ddot{\mathbf{R}}_{t,12})\boldsymbol{\Omega}_t^{(2)} \\
&= (\boldsymbol{\Omega}_t^{(2)})^\top(\boldsymbol{\Omega}_t^{(2)}\ddot{\mathbf{R}}_t(\boldsymbol{\Omega}_t^{(2)})^\top - \boldsymbol{\Omega}_t^{(2)}\ddot{\mathbf{R}}_t(\boldsymbol{\Omega}_t^{(1)})^\top(\ddot{\mathbf{R}}_{t,11})^{-1}\boldsymbol{\Omega}_t^{(1)}\ddot{\mathbf{R}}_t(\boldsymbol{\Omega}_t^{(2)})^\top)\boldsymbol{\Omega}_t^{(2)} \\
&= \mathbf{I}_t^{(2)}(\ddot{\mathbf{R}}_t - \ddot{\mathbf{R}}_t(\boldsymbol{\Omega}_t^{(1)})^\top(\ddot{\mathbf{R}}_{t,11})^{-1}\boldsymbol{\Omega}_t^{(1)}\ddot{\mathbf{R}}_t)\mathbf{I}_t^{(2)} \\
&= \mathbf{I}_t^{(2)}\nabla_t\ddot{\mathbf{R}}_t\mathbf{I}_t^{(2)}
\end{aligned} \tag{160}$$

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

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

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

we are breaking the multivariate expectation into an inner and outer expectation.

$$\begin{aligned}
\widetilde{\mathbf{O}}_t &= \mathrm{E}_{XY}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)] = \mathrm{E}_X[\,\mathrm{E}_{Y|x}[\boldsymbol{Y}_t\boldsymbol{Y}_t^\top|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]] \\
&= \mathrm{E}_X\big[\,\mathrm{var}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t] \\
&\quad + \mathrm{E}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]\,\mathrm{E}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t=\boldsymbol{x}_t]^\top\big] \\
&= \mathrm{E}_X[\mathbf{I}_t^{(2)}\nabla_t\ddot{\mathbf{R}}_t\mathbf{I}_t^{(2)}] + \mathrm{E}_X[(\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}_t\boldsymbol{X}_t-\mathbf{a}_t))(\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}_t\boldsymbol{X}_t-\mathbf{a}_t))^\top] \\
&= \mathbf{I}_t^{(2)}\nabla_t\ddot{\mathbf{R}}_t\mathbf{I}_t^{(2)} + \mathrm{var}_X\big[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}_t\boldsymbol{X}_t-\mathbf{a}_t)\big] \\
&\quad + \mathrm{E}_X[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}_t\boldsymbol{X}_t-\mathbf{a}_t)]\,\mathrm{E}_X[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}_t\boldsymbol{X}_t-\mathbf{a}_t)]^\top \\
&= \mathbf{I}_t^{(2)}\nabla_t\ddot{\mathbf{R}}_t\mathbf{I}_t^{(2)} + \mathbf{I}_t^{(2)}\nabla_t\mathbf{Z}_t\widetilde{\boldsymbol{V}}_t\mathbf{Z}_t^\top\nabla_t^\top\mathbf{I}_t^{(2)} + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{y}}_t^\top
\end{aligned} \tag{161}$$

Thus,

$$\widetilde{\mathbf{O}}_t = \mathbf{I}_t^{(2)}(\nabla_t\ddot{\mathbf{R}}_t + \nabla_t\mathbf{Z}_t\widetilde{\boldsymbol{V}}_t\mathbf{Z}_t^\top\nabla_t^\top)\mathbf{I}_t^{(2)} + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{y}}_t^\top \tag{162}$$

and

$$\mathrm{var}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1)] = \mathbf{I}_t^{(2)}(\nabla_t\ddot{\mathbf{R}}_t + \nabla_t\mathbf{Z}_t\widetilde{\boldsymbol{V}}_t\mathbf{Z}_t^\top\nabla_t^\top)\mathbf{I}_t^{(2)} \tag{163}$$

The variance can be decomposed into two parts via the law of total variance:

$$\mathrm{var}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1)] = \mathrm{E}_X[\,\mathrm{var}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t]] + \mathrm{var}_X[\,\mathrm{E}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t]] \tag{164}$$

Using equations 164, 160, and 163, we can solve for the variance (over $\boldsymbol{x}_t$) of the expected value of $\boldsymbol{Y}_t|\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1)$ conditioned on $\boldsymbol{X}_t = \boldsymbol{x}_t$:

$$\begin{aligned}
&\mathrm{var}_X[\,\mathrm{E}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t]] \\
&\qquad = \mathrm{var}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1)] - \mathrm{E}_X[\,\mathrm{var}_{Y|x}[\boldsymbol{Y}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1),\boldsymbol{X}_t]] \\
&\qquad = \mathbf{I}_t^{(2)}(\nabla_t\ddot{\mathbf{R}}_t + \nabla_t\mathbf{Z}_t\widetilde{\boldsymbol{V}}_t\mathbf{Z}_t^\top\nabla_t^\top)\mathbf{I}_t^{(2)} - \mathbf{I}_t^{(2)}\nabla_t\ddot{\mathbf{R}}_t\mathbf{I}_t^{(2)} \\
&\qquad = \mathbf{I}_t^{(2)}\nabla_t\mathbf{Z}_t\widetilde{\boldsymbol{V}}_t\mathbf{Z}_t^\top\nabla_t^\top\mathbf{I}_t^{(2)}
\end{aligned} \tag{165}$$

Though this variance is not used in the EM algoritm, it gives us the confidence intervals for the expected value of missing data while the variance of $\boldsymbol{Y}_t|\boldsymbol{Y}_t(1) = \boldsymbol{y}_t(1)$ gives us the prediction intervals for missing data.

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

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

$$\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)] = \boldsymbol{y}_t\,\mathrm{E}[\boldsymbol{X}_t^\top] = \boldsymbol{y}_t\widetilde{\mathbf{x}}_t^\top \tag{166}$$

If there are no observed values at time t, then

$$\begin{aligned}
&\mathrm{E}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)] \\
&\qquad = \mathrm{E}[(\mathbf{Z}_t\boldsymbol{X}_t+\mathbf{a}_t+\boldsymbol{V}_t)\boldsymbol{X}_t^\top] \\
&\qquad = \mathrm{E}[\mathbf{Z}_t\boldsymbol{X}_t\boldsymbol{X}_t^\top + \mathbf{a}_t\boldsymbol{X}_t^\top + \boldsymbol{V}_t\boldsymbol{X}_t^\top] \\
&\qquad = \mathbf{Z}_t\widetilde{\mathbf{P}}_t + \mathbf{a}_t\widetilde{\mathbf{x}}_t^\top + \mathrm{cov}[\boldsymbol{V}_t,\boldsymbol{X}_t] + \mathrm{E}[\boldsymbol{V}_t]\,\mathrm{E}[\boldsymbol{X}_t]^\top \\
&\qquad = \mathbf{Z}_t\widetilde{\mathbf{P}}_t + \mathbf{a}_t\widetilde{\mathbf{x}}_t^\top
\end{aligned} \tag{167}$$

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

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

$$\begin{aligned}
\widetilde{\mathbf{y}\mathbf{x}}_t &= \mathrm{E}_{XY}[\boldsymbol{Y}_t\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)] \\
&= \mathrm{cov}[\boldsymbol{Y}_t,\boldsymbol{X}_t|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1)] + \mathrm{E}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]\,\mathrm{E}_{XY}[\boldsymbol{X}_t^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]^\top \\
&= \mathrm{cov}[\boldsymbol{y}_t-\nabla_t(\boldsymbol{y}_t-\mathbf{Z}_t\boldsymbol{X}_t-\mathbf{a}_t)+\boldsymbol{V}_t^*,\boldsymbol{X}_t] + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_t^\top \\
&= \mathrm{cov}[\boldsymbol{y}_t,\boldsymbol{X}_t] - \mathrm{cov}[\nabla_t\boldsymbol{y}_t,\boldsymbol{X}_t] + \mathrm{cov}[\nabla_t\mathbf{Z}_t\boldsymbol{X}_t,\boldsymbol{X}_t] + \mathrm{cov}[\nabla_t\mathbf{a}_t,\boldsymbol{X}_t] \\
&\quad + \mathrm{cov}[\boldsymbol{V}_t^*,\boldsymbol{X}_t] + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_t^\top \\
&= 0 - 0 + \nabla_t\mathbf{Z}_t\widetilde{\boldsymbol{V}}_t + 0 + 0 + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_t^\top \\
&= \nabla_t\mathbf{Z}_t\widetilde{\boldsymbol{V}}_t + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_t^\top
\end{aligned} \tag{168}$$

This uses the computational formula for covariance: $\mathrm{E}[\boldsymbol{Y}\boldsymbol{X}^\top] = \mathrm{cov}[\boldsymbol{Y},\boldsymbol{X}] + \mathrm{E}[\boldsymbol{Y}]\,\mathrm{E}[\boldsymbol{X}]^\top$. $\boldsymbol{V}_t^*$ is a random variable with mean 0 and variance $\ddot{\mathbf{R}}_{t,22} - \ddot{\mathbf{R}}_{t,21}(\ddot{\mathbf{R}}_{t,11})^{-1}\ddot{\mathbf{R}}_{t,12}$ from equation 156. $\boldsymbol{V}_t^*$ and $\boldsymbol{X}_t$ are independent of each other, thus $\mathrm{cov}[\boldsymbol{V}_t^*, \boldsymbol{X}_t^\top] = 0$.

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

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

$$\begin{aligned}
\widetilde{\mathbf{y}\mathbf{x}}_t &= \mathrm{E}_{XY}[\boldsymbol{Y}_t\boldsymbol{X}_{t-1}^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)] \\
&= \mathrm{cov}[\boldsymbol{Y}_t,\boldsymbol{X}_{t-1}|\boldsymbol{Y}_t(1)=\boldsymbol{y}_t(1)] + \mathrm{E}_{XY}[\boldsymbol{Y}_t|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]\,\mathrm{E}_{XY}[\boldsymbol{X}_{t-1}^\top|\boldsymbol{Y}(1)=\boldsymbol{y}(1)]^\top \\
&= \mathrm{cov}[\boldsymbol{y}_t - \nabla_t(\boldsymbol{y}_t - \mathbf{Z}_t\boldsymbol{X}_t - \mathbf{a}_t) + \boldsymbol{V}_t^*, \boldsymbol{X}_{t-1}] + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_{t-1}^\top \\
&= \mathrm{cov}[\boldsymbol{y}_t,\boldsymbol{X}_{t-1}] - \mathrm{cov}[\nabla_t\boldsymbol{y}_t,\boldsymbol{X}_{t-1}] + \mathrm{cov}[\nabla_t\mathbf{Z}_t\boldsymbol{X}_t,\boldsymbol{X}_{t-1}] \\
&\quad + \mathrm{cov}[\nabla_t\mathbf{a}_t,\boldsymbol{X}_{t-1}] + \mathrm{cov}[\boldsymbol{V}_t^*,\boldsymbol{X}_{t-1}] + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_{t-1}^\top \\
&= 0 - 0 + \nabla_t\mathbf{Z}_t\widetilde{\boldsymbol{V}}_{t,t-1} + 0 + 0 + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_{t-1}^\top \\
&= \nabla_t\mathbf{Z}_t\widetilde{\boldsymbol{V}}_{t,t-1} + \widetilde{\mathbf{y}}_t\widetilde{\mathbf{x}}_{t-1}^\top
\end{aligned} \tag{169}$$

7 Degenerate variance models

It is possible that the model has deterministic and stochastic elements; mathematically this means that either $\mathbf{G}_t$, $\mathbf{H}_t$ or $\mathbf{F}$ have all zero rows, and this means that some of the observation or state processes are deterministic[17] Such models often arise when a MAR-p is put into MARSS-1 form. Assuming the model is solvable (one solution and not over-determined), we can modify the Kalman smoother and EM algorithm to handle models with deterministic elements.

The motivation behind the degenerate variance modification is that we want to use one set of EM update equations for all models in the MARSS class—regardless of whether they are partially or fully degenerate[18]. The difficulties arise in getting the $\mathbf{u}$ and $\boldsymbol{\xi}$ update equations. If we were to fix these or make $\boldsymbol{\xi}$ stochastic (a fixed mean and fixed variance), most of the trouble in this section could be avoided. However, fixing $\boldsymbol{\xi}$ or making it stochastic is putting a prior on it and placing a prior on the variance-covariance structure of $\boldsymbol{\xi}$ that conflicts logically with the model is often both unavoidable (since the correct variance-covariance structure depends on the parameters you are trying to estimate) and disastrous to one's estimation although the problem is often difficult to detect especially with long time series. Many papers have commented on this subtle problem. So, we want to be able to estimate $\boldsymbol{\xi}$ so we do not have to specify $\boldsymbol{\Lambda}$ (because we remove it from the model altogether). Note that in a univariate $\boldsymbol{x}$ model (one state), $\boldsymbol{\Lambda}$ is just a variance so we do not run into this trouble. The problems arise when $\boldsymbol{x}$ is multivariate (>1 state) and then we have to deal with the variance-covariance structure of the initial states.

7.1 Rewriting the state and observation models for degenerate variance systems

Let's start with an example where $y_{2,t}$ (2nd y) has no added observation error.

$$\mathbf{R}_t = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix} \text{ and } \mathbf{H}_t = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \tag{170}$$

Let $\boldsymbol{\Omega}_{t,r}^+$ be a $p \times n$ matrix that extracts the p non-zero rows from $\mathbf{H}_t$. The diagonal matrix $(\boldsymbol{\Omega}_{t,r}^+)^\top\boldsymbol{\Omega}_{t,r}^+ \equiv \mathbf{I}_{t,r}^+$ is a diagonal matrix that can zero out the $\mathbf{H}_t$ zero rows in any n row matrix.

[17] Deterministic means that given the parameters, the states or observation processes have known values and are not random variables.

[18] Degenerate means zeros on the diagonal of the variance-covariance matrix, which appears as a zero row in $\mathbf{G}_t$, $\mathbf{H}_t$ or $\mathbf{F}$.

$$\begin{aligned}
\boldsymbol{\Omega}_{t,r}^{+} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{I}_{t,r}^{+} = (\boldsymbol{\Omega}_{t,r}^{+})^{\top}\boldsymbol{\Omega}_{t,r}^{+} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
\boldsymbol{y}_t^{+} &= \boldsymbol{\Omega}_{t,r}^{+}\boldsymbol{y}_t = \begin{bmatrix} y_1 \\ y_3 \end{bmatrix}_t \qquad \boldsymbol{y}_t^{+} = \mathbf{I}_{t,r}^{+}\boldsymbol{y}_t = \begin{bmatrix} y_1 \\ 0 \\ y_3 \end{bmatrix}_t
\end{aligned} \tag{171}$$

Let $\boldsymbol{\Omega}_{t,r}^{(0)}$ be a $(n-p)\times n$ matrix that extracts the $n-p$ zero rows from $\mathbf{H}_t$. For the example above,

$$\begin{aligned}
\boldsymbol{\Omega}_{t,r}^{(0)} &= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \qquad \mathbf{I}_{t,r}^{(0)} = (\boldsymbol{\Omega}_{t,r}^{(0)})^{\top}\boldsymbol{\Omega}_{t,r}^{(0)} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
\boldsymbol{y}_t^{(0)} &= \boldsymbol{\Omega}_{t,r}^{(0)}\boldsymbol{y}_t = \begin{bmatrix} y_3 \end{bmatrix}_t \qquad \boldsymbol{y}_t^{(0)} = \mathbf{I}_{t,r}^{(0)}\boldsymbol{y}_t = \begin{bmatrix} 0 \\ y_2 \\ 0 \end{bmatrix}_t
\end{aligned} \tag{172}$$

Similarly, $\boldsymbol{\Omega}_{t,q}^{+}$ extracts the states associated with the non-zero rows in $\mathbf{G}_t$ and $\boldsymbol{\Omega}_{t,q}^{(0)}$ extracts the zero rows. $\mathbf{I}_{t,q}^{+}$ and $\mathbf{I}_{t,q}^{(0)}$ are defined similarly.

Using these definitions, we can rewrite the state process part of the MARSS model by separating out the deterministic parts. $\boldsymbol{x}_t^{(0)}$ is the rows of $\boldsymbol{x}_t$ that are associated with all-zero rows of $\mathbf{G}_t$, that means there is no w_t in the x_t equation for those rows[19]

$$\begin{aligned}
\boldsymbol{x}_t^{(0)} &= \boldsymbol{\Omega}_{t,q}^{(0)}\boldsymbol{x}_t = \boldsymbol{\Omega}_{t,q}^{(0)}(\mathbf{B}_t\boldsymbol{x}_{t-1} + \mathbf{u}_t) \\
\boldsymbol{x}_t^{+} &= \boldsymbol{\Omega}_{t,q}^{+}\boldsymbol{x}_t = \boldsymbol{\Omega}_{t,q}^{+}(\mathbf{B}_t\boldsymbol{x}_{t-1} + \mathbf{u}_t + \mathbf{G}_t\boldsymbol{w}_t) \\
\boldsymbol{w}_t^{+} &\sim \text{MVN}(0, \mathbf{Q}_t)
\end{aligned} \tag{173}$$

Similarly, we can rewrite the observation process part of the MARSS model by separating out the parts with no observation error:

$$\begin{aligned}
\boldsymbol{y}_t^{(0)} &= \boldsymbol{\Omega}_{t,r}^{(0)}\boldsymbol{y}_t = \boldsymbol{\Omega}_{t,r}^{(0)}(\mathbf{Z}_t\boldsymbol{x}_t + \mathbf{a}_t) \\
&= \boldsymbol{\Omega}_{t,r}^{(0)}(\mathbf{Z}_t\mathbf{I}_{t,q}^{+}\boldsymbol{x}_t + \mathbf{Z}_t\mathbf{I}_{t,q}^{(0)}\boldsymbol{x}_t + \mathbf{a}_t) \\
\boldsymbol{y}_t^{+} &= \boldsymbol{\Omega}_{t,r}^{+}\boldsymbol{y}_t = \boldsymbol{\Omega}_{t,r}^{+}(\mathbf{Z}_t\boldsymbol{x}_t + \mathbf{a}_t + \mathbf{H}_t\boldsymbol{v}_t) \\
&= \boldsymbol{\Omega}_{t,r}^{+}(\mathbf{Z}_t\mathbf{I}_{t,q}^{+}\boldsymbol{x}_t + \mathbf{Z}_t\mathbf{I}_{t,q}^{(0)}\boldsymbol{x}_t + \mathbf{a}_t + \mathbf{H}_t\boldsymbol{v}_t) \\
\boldsymbol{v}_t^{+} &\sim \text{MVN}(0, \mathbf{R}_t)
\end{aligned} \tag{174}$$

In order for this to be solvable using an EM algorithm with the Kalman filter, we require that no estimated $\mathbf{B}$ or $\mathbf{u}$ elements appear in the equation for $\boldsymbol{y}_t^{(0)}$ (via x_t in that equation). Since the $\boldsymbol{y}_t^{(0)}$ do not appear in the likelihood function (since $\mathbf{H}_t^{(0)} = 0$), $\boldsymbol{y}_t^{(0)}$ would not affect the estimate for the parameters appearing in the $\boldsymbol{y}_t^{(0)}$ equation. This translates to the following constraints, $(1_{1\times m} \otimes \boldsymbol{\Omega}_{t,r}^{(0)}\mathbf{Z}_t\mathbf{I}_{t,q}^{(0)})\mathbf{D}_{t,b}$ is all zeros and $\boldsymbol{\Omega}_{t,r}^{(0)}\mathbf{Z}_t\mathbf{I}_{t,q}^{(0)}\mathbf{D}_u$ is all zeros. Also notice that $\boldsymbol{\Omega}_{t,r}^{(0)}\mathbf{Z}_t$ and $\boldsymbol{\Omega}_{t,r}^{(0)}\mathbf{a}_t$ appear in the $\boldsymbol{y}^{(0)}$ equation and not in the $\boldsymbol{y}^{+}$ equation. This means that $\boldsymbol{\Omega}_{t,r}^{(0)}\mathbf{Z}_t$ and $\boldsymbol{\Omega}_{t,r}^{(0)}\mathbf{a}_t$ must be only fixed terms.

In summary, the degenerate model can be reduced to the following (with $\boldsymbol{x}_0$ not specified yet).

$$\begin{aligned}
\boldsymbol{x}_t^{(0)} &= \mathbf{B}_t^{(0)}\boldsymbol{x}_{t-1} + \mathbf{u}_t^{(0)} \\
\boldsymbol{x}_t^{+} &= \mathbf{B}_t^{+}\boldsymbol{x}_{t-1} + \mathbf{u}_t^{+} + \mathbf{G}_t^{+}\boldsymbol{w}_t \\
\boldsymbol{w}_t &\sim \text{MVN}(0, \mathbf{Q}_t) \\
\boldsymbol{y}_t^{(0)} &= \mathbf{Z}^{(0)}\mathbf{I}_q^{+}\boldsymbol{x}_t + \mathbf{Z}^{(0)}\mathbf{I}_q^{(0)}\boldsymbol{x}_t + \mathbf{a}_t^{(0)} \\
\boldsymbol{y}_t^{+} &= \mathbf{Z}_t^{+}\boldsymbol{x}_t + \mathbf{a}_t^{+}\mathbf{H}_t^{+}\boldsymbol{v}_t \\
&= \mathbf{Z}_t^{+}\mathbf{I}_q^{+}\boldsymbol{x}_t + \mathbf{Z}_t^{+}\mathbf{I}_q^{(0)}\boldsymbol{x}_t + \mathbf{a}_t^{+} + \mathbf{H}_t^{+}\boldsymbol{v}_t \\
\boldsymbol{v}_t &\sim \text{MVN}(0, \mathbf{R})
\end{aligned} \tag{175}$$

[19] $x_{t,i} = \mathbf{B}_{t,i}\boldsymbol{x}_{t-1} + \mathbf{u}_{t,i}$ where the i subscript means i-th row.

where $\mathbf{B}_t^{(0)} = \mathbf{\Omega}_{t,q}^{(0)}\mathbf{B}_t$ and $\mathbf{B}_t^{+} = \mathbf{\Omega}_{t,q}^{+}\mathbf{B}_t$ so that $\mathbf{B}_t^{(0)}$ are the rows of $\mathbf{B}_t$ corresponding to the zero rows of $\mathbf{G}_t$ and $\mathbf{B}_t^{+}$ are the rows of $\mathbf{B}_t$ corresponding to non-zero rows of $\mathbf{G}_t$. The other parameters are similarly defined: $\mathbf{u}_t^{(0)} = \mathbf{\Omega}_{t,q}^{(0)}\mathbf{u}_t$ and $\mathbf{u}_t^{+} = \mathbf{\Omega}_{t,q}^{+}\mathbf{u}_t$, $\mathbf{Z}_t^{(0)} = \mathbf{\Omega}_{t,r}^{(0)}\mathbf{Z}_t$ and $\mathbf{Z}_t^{+} = \mathbf{\Omega}_{t,r}^{+}\mathbf{Z}_t$, and $\mathbf{a}_t^{(0)} = \mathbf{\Omega}_{t,r}^{(0)}\mathbf{a}_t$ and $\mathbf{a}_t^{+} = \mathbf{\Omega}_{t,r}^{+}\mathbf{a}_t$.

7.2 Identifying the fully deterministic $\boldsymbol{x}$ rows

To derive EM update equations, we need to take the derivative of the expected log-likelihood holding everything but the parameter of interest constant. If there are deterministic $\boldsymbol{x}_t$ rows, then we cannot hold these constant and do this partial differentiation with respect to the state parameters. We need to identify these $\boldsymbol{x}_t$ rows and remove them from the likelihood function by rewriting them in terms of only the state parameters[20]. For this derivation, I am going to make the simplifying assumption that the locations of the all-zero rows in $\mathbf{G}_t$ and $\mathbf{H}_t$ are time-invariant. This is not strictly necessary, but simplifies the algebra greatly.

For the deterministic $\boldsymbol{x}_t$ rows, denoted $\boldsymbol{x}_t^d$, the process equation is $\boldsymbol{x}_t = \mathbf{B}_t\boldsymbol{x}_{t-1} + \mathbf{u}_t$, with no $\boldsymbol{w}_t$ term. When we do the partial differentiation step in deriving the EM update equation for $\mathbf{u}$, $\mathbf{B}$ or $\boldsymbol{\xi}$, we will need to take a partial derivative while holding $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ constant. We cannot hold the deterministic rows of $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ constant while changing the corresponding rows of $\mathbf{u}_t$ and $\mathbf{B}_t$ (or $\boldsymbol{\xi}$ if $t = 0$ or $t = 1$). If a row of $\boldsymbol{x}_t$ is fully deterministic, then that $x_{i,t}$ must change when row i of $\mathbf{u}_t$ or $\mathbf{B}_t$ is changed. Thus we cannot do the partial differentiation step required in the EM update equation derivation.

So we need to identify the fully deterministic $\boldsymbol{x}_t$ and treat them differently in our likelihood so we can derive the update equation. First I will define some terminology regarding the $\boldsymbol{x}_t$.

- (0) rows of any $\boldsymbol{x}$, $\mathbf{B}$, $\mathbf{u}$ or $\mathbf{I}$ matrix that are associated with all-zero rows of $\mathbf{G}_t$, e.g. $\boldsymbol{x}_t^{(0)}$.
- (+) rows of any $\boldsymbol{x}$, $\mathbf{B}$, $\mathbf{u}$ or $\mathbf{I}$ matrix that are associated with non-zero rows of $\mathbf{G}_t$, e.g. $\boldsymbol{x}_t^{(+)}$.
- 'directly stochastic' $\boldsymbol{x}_t$ are denoted $\boldsymbol{x}_t^{ds}$. These are the same as $\boldsymbol{x}_t^{+}$. These $\boldsymbol{x}_t$ have a w_t from their row of $\mathbf{G}_t$.
- 'deterministic' $\boldsymbol{x}_t$ are denoted $\boldsymbol{x}_t^{d}$. These are those $\boldsymbol{x}_t^{(0)}$ which have no w_t terms either from their own row or picked up through $\mathbf{B}$ from a non-zero row of $\mathbf{G}_t$.
- 'indirectly stochastic' $\boldsymbol{x}_t$ are denoted $\boldsymbol{x}_t^{is}$. Indirectly stochastic $\boldsymbol{x}_t^{is}$ have a corresponding row of $\mathbf{G}_t$ that is all zero, but pick up a w_t from a non-zero row of $\mathbf{G}_t$ through $\mathbf{B}$ in one of the prior $\mathbf{B}_t\boldsymbol{x}_t$ steps.

The stochastic $\boldsymbol{x}_t$ are denoted $\boldsymbol{x}_t^{s}$ whether they are indirectly or directly stochastic.

How do you determine the d, or deterministic, set of $\boldsymbol{x}_t$ rows? These are the rows of $\boldsymbol{x}_t$ with no w terms, from time t or from prior t. Note that the location of the d rows is time-dependent, a row may be deterministic at time t but pick up a w at time $t+1$ and thus be indirectly stochastic thereafter. I am requiring that once a row becomes indirectly stochastic, it remains stochastic; rows are not allowed to flip back and forth between deterministic (no w terms in them) and stochastic (containing a w term).

I will work through an example and then show a general algorithm to keep track of the deterministic rows at time t.

Example

Let $\boldsymbol{x}_0 = \boldsymbol{\xi}$ (so $\mathbf{F}$ is all zero and $\boldsymbol{x}_0$ is not stochastic). Define $\mathbf{I}_t^{ds}$, $\mathbf{I}_t^{is}$, and $\mathbf{I}_t^{d}$ as diagonal indicator matrices with a 1 at $\mathbf{I}(i,i)$ if row i is directly stochastic, indirectly stochastic, or deterministic respectively. $\mathbf{I}_t^{ds}+\mathbf{I}_t^{is}+\mathbf{I}_t^{d} = \mathbf{I}_m$. Let our state equation be $\boldsymbol{x}_t = \mathbf{B}\boldsymbol{x}_{t-1} + \mathbf{G}w_t$. Let

$$\mathbf{B} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{G} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{176}$$

At $t = 0$, $\boldsymbol{x}_0$ is fixed, aka deterministic.

$$\boldsymbol{x}_0 = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{bmatrix} \tag{177}$$

[20] Then we can do the partial differentiation with respect to the parameters.

$$\mathbf{I}_0^d = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{I}_0^s = \mathbf{I}_0^{is} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{178}$$

At $t = 1$, the $\boldsymbol{x}_t$ begin picking up w_t starting with $x_{1,t}$.

$$\boldsymbol{x}_1 = \begin{bmatrix} \pi_1 + \pi_2 + w_1 \\ \pi_1 \\ \pi_2 \\ \pi_4 \end{bmatrix} \tag{179}$$

$$\mathbf{I}_1^d = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{I}_1^{ds} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{I}_1^{is} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{180}$$

At $t = 2$, $x_{2,2}$ picks up w_1 through $\mathbf{B}$.

$$\boldsymbol{x}_2 = \begin{bmatrix} \cdots + w_2 \\ \pi_1 + \pi_2 + w_1 \\ \pi_1 \\ \pi_4 \end{bmatrix} \tag{181}$$

$$\mathbf{I}_2^d = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{I}_2^{ds} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{I}_0^{is} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{182}$$

By $t = 3$, the $\mathbf{I}^d$ and $\mathbf{I}^{is}$ stabilize.

$$\boldsymbol{x}_3 = \begin{bmatrix} \cdots + w_1 + w_2 + w_3 \\ \cdots + w_1 + w_2 \\ \pi_1 + \pi_2 + w_1 \\ \pi_4 \end{bmatrix} \tag{183}$$

$$\mathbf{I}_3^d = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{I}_3^{ds} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{I}_3^{is} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{184}$$

After time $t = 3$ the location of the deterministic and indirectly stochastic rows is stabilized and no longer changes.

Finding the indirectly stochastic rows

In general, it can take up to m time steps for the location of the deterministic rows to stabilize. This is because $\mathbf{B}_t$ is like an adjacency matrix, and I require that the location of the 0 elements in $\mathbf{B}_1\mathbf{B}_2 \dots \mathbf{B}_t$ is time invariant. If we replace all non-zero elements in $\mathbf{B}_t$ with 1, then we have an adjacency matrix, let's call it $\mathbf{M}$. If there is a path in $\mathbf{M}$ from $x_{j,t}$, where j is a (0) row of $\boldsymbol{x}$, to an $x_{i,t}$, where i is a (+) row, then row j of $\boldsymbol{x}$ will eventually be indirectly stochastic. Graph theory tells us that it takes at most m steps for a $m \times m$ adjacency matrix to show full connectivity. This means that if element (j, i) is 0 in M^m then row j is not connected to row i by any path and thus will remain unconnected for $M^{t>m}$; note element i, j can be 0 while j, i is not.

This means that to determine if $x_{j,t}$, in the (0) rows, is indirectly stochastic, we raise $\mathbf{M}$, to the t power and look if there is a non-zero value in the j-th row and any (+) columns of $\mathbf{M}^t$. In words, we looking for a path from $x_{j,t}$ to any x_+ in the past. We do not need to do this past $t = m$ since the location of the indirectly stochastic and deterministic rows stabilize by then.

Since my $\mathbf{B}_t$ matrices are small, I use an inefficient strategy in the MARSS code to construct the indicator matrices $\mathbf{I}_d^t$. I define $\mathbf{M}$ as $\mathbf{B}_t$ with the non-zero $\mathbf{B}$ replaced with 1; I require that the location of the non-zero elements in $\mathbf{B}_t$ are time-invariant so there is only one $\mathbf{M}$. Within the product $\mathbf{M}^t$, those rows where only 0s appear in the 'stochastic' columns (non-zero $\mathbf{G}_t$ rows) are the fully deterministic $\boldsymbol{x}_{t+1}$ rows. Note, $t + 1$ so

one time step ahead. There are much faster algorithms for finding paths, but my $\mathbf{M}$ tend to be small. Also, unfortunately, using $\mathbf{B}^t$ in place of $\mathbf{M}^t$ is not robust. Let's say $\mathbf{B} = \left[\begin{smallmatrix} -1 & -2 \\ 1 & 1 \end{smallmatrix}\right]$, $\mathbf{G} = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ and $\boldsymbol{x}_0$ is fixed (not stochastic). $\mathbf{B}^2$ is a diagonal matrix suggesting that no connection between x_2 and x_1 at time $t = 2$. That is incorrect. $x_{2,t}$ is indirectly stochastic.

7.2.1 Redefining the $\boldsymbol{x}_t^d$ elements in the likelihood

Because the deterministic rows of $\boldsymbol{x}_t$ do not appear in the $\boldsymbol{x}$ part of the likelihood (no error term = no likelihood), we have to move them into the $\boldsymbol{y}$ part of the likelihood. To do that we need to re-write them in terms of only model parameters and remove all $\boldsymbol{x}_{t-1}$ terms. This section walks through how to do that.

By definition, all the $\mathbf{B}_t$ elements in the ds and is columns of the d rows of $\mathbf{B}_t$ are 0. If they weren't, then $\boldsymbol{x}_t^d$ wouldn't be a deterministic row because it would pick up a w from a directly or indirectly stochastic x from a prior $t-1$. This is due to the constraint that I have imposed that locations of 0s in $\mathbf{B}_t$ are time-invariant and the location of the zero rows in $\mathbf{G}_t$ also time-invariant: $\mathbf{I}_q^+$ and $\mathbf{I}_q^{(0)}$ are time-constant.

Example

Consider this $\mathbf{B}$ and $\mathbf{G}$, which would arise in a MARSS version of an AR-3 model:

$$\mathbf{B} = \begin{bmatrix} b_1 & b_2 & b_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \mathbf{G} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \tag{185}$$

Using $\boldsymbol{x}_0 = \boldsymbol{\xi}$ (so fixed and not stochastic):

$$\boldsymbol{x}_0 = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{bmatrix} \quad \boldsymbol{x}_1 = \begin{bmatrix} \cdots + w_1 \\ \pi_1 \\ \pi_2 \end{bmatrix} \quad \boldsymbol{x}_2 = \begin{bmatrix} \cdots + w_2 \\ \cdots + w_1 \\ \pi_1 \end{bmatrix} \quad \boldsymbol{x}_3 = \begin{bmatrix} \cdots + w_3 \\ \cdots + w_2 \\ \cdots + w_1 \end{bmatrix} \tag{186}$$

The ... just represent 'some values'. The key part is the w appearing which is the stochasticity. At $t = 1$, rows 2 and 3 are deterministic. At $t = 2$, row 3 is deterministic, and at $t = 3$, no rows are deterministic.

The $\mathbf{I}^d$ are:

$$\mathbf{I}_{q,1}^d = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{I}_{q,2}^d = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{I}_{q,3}^d = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{187}$$

The $\mathbf{M}$ are:

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \mathbf{M}^2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{I}_{q,3}^d = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{188}$$

We can rewrite the equation for the deterministic rows in $\boldsymbol{x}_t$ as follows. $\boldsymbol{x}_t^d$ is $\boldsymbol{x}_t$ with the d rows zeroed out, so $\boldsymbol{x}_t^d = \mathbf{I}_{q,t}^d \boldsymbol{x}_t$.

$$\begin{aligned} \boldsymbol{x}_1^d &= \mathbf{I}_{q,1}^d \boldsymbol{x}_1 \\ &= \mathbf{I}_{q,1}^d (\mathbf{B}_1 \boldsymbol{x}_0 + \mathbf{f}_{u,1} + \mathbf{D}_{u,1} \boldsymbol{v}) \\ \boldsymbol{x}_2^d &= \mathbf{I}_{q,2}^d \boldsymbol{x}_2 \\ &= \mathbf{I}_{q,2}^d (\mathbf{B}_2 \boldsymbol{x}_1 + \mathbf{u}_2) \\ &= \mathbf{I}_{q,2} \mathbf{B}_2 ((\mathbf{B}_1 \boldsymbol{x}_0 + \mathbf{f}_{u,1} + \mathbf{D}_{u,1} \boldsymbol{v}) + \mathbf{f}_{u,2} + \mathbf{D}_{u,2} \boldsymbol{v}) \\ &= \mathbf{I}_2^d (\mathbf{B}_2 \mathbf{B}_1 \boldsymbol{x}_0 + \mathbf{B}_2 \mathbf{f}_{1,u} + \mathbf{f}_{2,u} + (\mathbf{B}_2 \mathbf{D}_{u,1} + \mathbf{D}_{u,2}) \boldsymbol{v}) \\ &\dots \end{aligned} \tag{189}$$

The messy part is keeping track of which rows are deterministic because this will potentially change up to time $t = m$.

We can rewrite the function for $\boldsymbol{x}_t^d$, where t_0 is the t at which the initial state is defined. It is either $t = 0$ or $t = 1$.

$$\begin{aligned}
\boldsymbol{x}_t^d &= \mathbf{I}_t^d(\mathbf{B}_t^*\boldsymbol{x}_{t_0} + \mathbf{f}_t^* + \mathbf{D}_t^*\boldsymbol{v}) \\
\text{where} & \\
\mathbf{B}_{t_0}^* &= \mathbf{I}_m \\
\mathbf{B}_t^* &= \mathbf{B}_t\mathbf{B}_{t-1}^* \quad t > t_0 \\
& \\
\mathbf{f}_{t_0}^* &= 0 \\
\mathbf{f}_t^* &= \mathbf{B}_t\mathbf{f}_{t-1}^* + \mathbf{f}_{t,u} \quad t > t_0 \\
& \\
\mathbf{D}_{t_0}^* &= 0 \\
\mathbf{D}_t^* &= \mathbf{B}_t\mathbf{D}_{t-1}^* + \mathbf{D}_{t,u} \quad t > t_0 \\
& \\
\mathbf{I}_{q,t_0}^d &= \mathbf{I}_\lambda^d \\
\text{diag}(\mathbf{I}_{t_0+\tau}^d) &= \text{apply}(\boldsymbol{\Omega}_q^{(0)}\mathbf{M}^\tau\boldsymbol{\Omega}_q^+ == 0, 1, \text{all})
\end{aligned} \tag{190}$$

The bottom line is written in R: $\mathbf{I}_{t_0+\tau}^d$ is a diagonal matrix with a 1 at (i, i) where row i of $\mathbf{G}$ is all 0 and all ds and is columns in row i of $\mathbf{M}^t$ are equal to zero.

In the expected log-likelihood, the term $\mathrm{E}[\boldsymbol{X}_t^d] = \mathrm{E}[\boldsymbol{X}_t^d|\boldsymbol{Y} = \boldsymbol{y}]$, meaning the expected value of $\boldsymbol{X}_t^d$ conditioned on the data, appears. Thus in the expected log-likelihood the function will be written:

$$\begin{aligned}
\boldsymbol{X}_t^d &= \mathbf{I}_t^d(\mathbf{B}_t^*\boldsymbol{X}_{t_0} + \mathbf{f}_t^* + \mathbf{D}_t^*\boldsymbol{v}) \\
\mathrm{E}[\boldsymbol{X}_t^d] &= \mathbf{I}_t^d(\mathbf{B}_t^*\,\mathrm{E}[\boldsymbol{X}_{t_0}] + \mathbf{f}_t^* + \mathbf{D}_t^*\boldsymbol{v})
\end{aligned} \tag{191}$$

When the j-th row of $\mathbf{F}$ is all zero, meaning the j-th row of $\boldsymbol{x}_0$ is fixed to be ξ_j, then $\mathrm{E}[X_{t_0,j}] \equiv \xi_j$. This is the case where we treat $x_{t_0,j}$ as fixed and we either estimate or specify its value. If $\boldsymbol{x}_{t_0}$ is wholly treated as fixed, then $\mathrm{E}[\boldsymbol{X}_{t_0}] \equiv \boldsymbol{\xi}$ and $\boldsymbol{\Lambda}$ does not appear in the model at all. In the general case, where some $x_{t_0,j}$ are treated as fixed and some as stochastic, we can write $\mathrm{E}[\boldsymbol{X}_t^d]$ appearing in the expected log-likelihood as:

$$\mathrm{E}[\boldsymbol{X}_{t_0}] = (\mathbf{I}_m - \mathbf{I}_\lambda^{(0)})\,\mathrm{E}[\boldsymbol{X}_{t_0}] + \mathbf{I}_\lambda^{(0)}\boldsymbol{\xi} \tag{192}$$

$\mathbf{I}_\lambda^{(0)}$ is a diagonal indicator matrix with 1 at (j, j) if row j of $\mathbf{F}$ is all zero.

If $\mathbf{B}^{d,d}$ and $\mathbf{u}^d$ are time-constant, we could use the matrix geometric series:

$$\begin{aligned}
\boldsymbol{x}_t^d =& (\mathbf{B}^{d,d})^t\boldsymbol{x}_0^d + \sum_{i=0}^{t-1}(\mathbf{B}^{d,d})^i\mathbf{u}^d = (\mathbf{B}^{d,d})^t\boldsymbol{x}_0^d + (\mathbf{I} - \mathbf{B}^{d,d})^{-1}(\mathbf{I} - (\mathbf{B}^{d,d})^t)\mathbf{u}^d, \quad \text{if } \mathbf{B}^{d,d} \neq \mathbf{I} \\
& \boldsymbol{x}_0^d + \mathbf{u}^d, \quad \text{if } \mathbf{B}^{d,d} = \mathbf{I}
\end{aligned} \tag{193}$$

where $\mathbf{B}^{d,d}$ is the block of d's associated with the deterministic $\boldsymbol{x}_t$.

7.2.2 Dealing with the $\boldsymbol{x}_t^{is}$ elements in the likelihood and associated parameter rows

Although $\boldsymbol{w}_t^{is} = 0$, these terms are connected to the stochastic $\boldsymbol{x}$'s in earlier time steps though $\mathbf{B}$, thus all $\boldsymbol{x}_t^{is}$ are possible for a given $\mathbf{u}_t$, $\mathbf{B}_t$ or $\boldsymbol{\xi}$. However, all $\boldsymbol{x}_t^{is}$ are not possible conditioned on $\boldsymbol{x}_{t-1}$, so we are back in the position that we cannot both change $\boldsymbol{x}_t$ and change $\mathbf{u}_t$.

Recall that for the partial differentiation step in the EM algorithm, we need to be able to hold the $E[\boldsymbol{X}_t]$ appearing in the likelihood constant. We can deal with the deterministic $\boldsymbol{x}_t$ because they are not stochastic and do not have 'expected values'. They can be removed from the likelihood by rewriting $\boldsymbol{x}_t^d$ in terms of the model parameters. We cannot do that for $\boldsymbol{x}_t^{is}$ because these x are stochastic. There is no equation for them; all $\boldsymbol{x}^{is}$ are possible but some are more likely than others. We also cannot replace $\boldsymbol{x}_t^{is}$ with $\mathbf{B}_t^{is}E[\boldsymbol{X}_{t-1}] + \mathbf{u}_t^{is}$ to force $\mathbf{B}_t^{is}$ and $\mathbf{u}^{is}$ to appear in the $\boldsymbol{y}$ part of the likelihood. The reason is that $\mathrm{E}[\boldsymbol{X}_t]$ and $\mathrm{E}[\boldsymbol{X}_{t-1}]$ both appear in the likelihood and we cannot hold both constant (as we must for the partial differentiation) and

at the same time change $\mathbf{B}_t^{is}$ or $\mathbf{u}_t^{is}$ as we are doing when we differentiate with respect to $\mathbf{B}_t^{is}$ or $\mathbf{u})_t^{is}$. We cannot do that because $\boldsymbol{x}_t^{is}$ is constrained to equal $\mathbf{B}_t^{is}\boldsymbol{x}_{t-1} + \mathbf{u}_t^{is}$.

This effectively means that we cannot estimate $\mathbf{B}_t^{is}$ and $\mathbf{u}_t^{is}$ because we cannot rewrite $\boldsymbol{x}_t^{is}$ in terms of only the model parameters. This is specific to the EM algorithm because it is an iterative algorithm where the expected $\boldsymbol{X}_t$ are computed with fixed parameters and then the $\mathrm{E}[\boldsymbol{X}_t]$ are held fixed at their expected values while the parameters are updated. In my $\mathbf{B}$ update equation, I assume that $\mathbf{B}_t^{(0)}$ is fixed for all t. Thus I circumvent the problem altogether for $\mathbf{B}$. For $\mathbf{u}$, I assume that only the $\mathbf{u}^{is}$ elements are fixed.

7.3 Expected log-likelihood for degenerate models

The basic idea is to replace $\mathbf{I}_q^d\,\mathrm{E}[\boldsymbol{X}_t]$ with a deterministic function involving only the state parameters (and $\mathrm{E}[\boldsymbol{X}_{t_0}]$ if $\boldsymbol{X}_{t_0}$ is stochastic) . These appear in the $\boldsymbol{y}$ part of the likelihood in $\mathbf{Z}_t\boldsymbol{X}_t$ when the d columns of $\mathbf{Z}_t$ have non-zero values. They appear in the $\boldsymbol{x}$ part of the likelihood in $\mathbf{B}_t\boldsymbol{X}_{t-1}$ when the d columns of $\mathbf{B}_t$ have non-zero values. They do not appear in $\boldsymbol{X}_t$ in the $\boldsymbol{x}$ part of the likelihood because $\mathbb{Q}_t$ has all the non-s columns and rows zeroed out (non-s includes both d and is) and the element to the left of $\mathbb{Q}_t$ is a row vector and to the right, it is a column vector. Thus any $\boldsymbol{x}_t^d$ in $\boldsymbol{X}_t$ are being zeroed out by $\mathbb{Q}_t$.

The first step is to pull out the $\mathbf{I}_t^d\boldsymbol{X}_t$:

$$
\begin{aligned}
\Psi^+ = \mathrm{E}[\log \boldsymbol{L}(\boldsymbol{Y}^+, \boldsymbol{X}^+; \Theta)] = \mathrm{E}[&-\frac{1}{2}\sum_1^T \\
&(\boldsymbol{Y}_t - \mathbf{Z}_t(\mathbf{I}_m - \mathbf{I}_t^d)\boldsymbol{X}_t - \mathbf{Z}_t\mathbf{I}_t^d\boldsymbol{X}_t - \mathbf{a}_t)^\top \mathbb{R}_t \\
&(\boldsymbol{Y}_t - \mathbf{Z}_t(\mathbf{I}_m - \mathbf{I}_t^d)\boldsymbol{X}_t - \mathbf{Z}_t\mathbf{I}_t^d\boldsymbol{X}_t - \mathbf{a}_t) - \frac{1}{2}\sum_1^T \log|\mathbf{R}_t| \\
&- \frac{1}{2}\sum_{t_0+1}^T (\boldsymbol{X}_t - \mathbf{B}_t((\mathbf{I}_m - \mathbf{I}_{t-1}^d)\boldsymbol{X}_{t-1} + \mathbf{I}_{t-1}^d\boldsymbol{X}_{t-1}) - \mathbf{u}_t)^\top \mathbb{Q}_t \\
&(\boldsymbol{X}_t - \mathbf{B}_t((\mathbf{I}_m - \mathbf{I}_{t-1}^d)\boldsymbol{X}_{t-1} + \mathbf{I}_{t-1}^d\boldsymbol{X}_{t-1}) - \mathbf{u}_t) - \frac{1}{2}\sum_{t_0+1}^T \log|\mathbf{Q}_t| \\
&- \frac{1}{2}(\boldsymbol{X}_{t_0} - \boldsymbol{\xi})^\top \mathbb{L}(\boldsymbol{X}_{t_0} - \boldsymbol{\xi}) - \frac{1}{2}\log|\boldsymbol{\Lambda}| - \frac{n}{2}\log 2\pi
\end{aligned}
\tag{194}
$$

See section 7.2 for the definition of $\mathbf{I}_t^d$.

Next we replace $\mathbf{I}_q^d\boldsymbol{X}_t$ with equation 190. $\boldsymbol{X}_{t_0}$ will appear in this function instead of $\boldsymbol{x}_{t_0}$. I rewrite $\mathbf{u}_t$ as $\mathbf{f}_{u,t} + \mathbf{D}_{u,t}\boldsymbol{v}$. This gives us the expected log-likelihood:

$$
\begin{aligned}
\Psi^+ = \mathrm{E}[\log \boldsymbol{L}(\boldsymbol{Y}^+, \boldsymbol{X}^+; \Theta)] = \mathrm{E}[&-\frac{1}{2}\sum_1^T \\
&(\boldsymbol{Y}_t - \mathbf{Z}_t(\mathbf{I}_m - \mathbf{I}_t^d)\boldsymbol{X}_t - \mathbf{Z}_t\mathbf{I}_t^d(\mathbf{B}_t^*\boldsymbol{X}_{t_0} + \mathbf{f}_t^* + \mathbf{D}_t^*\boldsymbol{v}) - \mathbf{a}_t)^\top \mathbb{R}_t \\
&(\boldsymbol{Y}_t - \mathbf{Z}_t(\mathbf{I}_m - \mathbf{I}_t^d)\boldsymbol{X}_t - \mathbf{Z}_t\mathbf{I}_t^d(\mathbf{B}_t^*\boldsymbol{X}_{t_0} + \mathbf{f}_t^* + \mathbf{D}_t^*\boldsymbol{v}) - \mathbf{a}_t) - \frac{1}{2}\sum_1^T \log|\mathbf{R}_t| \\
&- \frac{1}{2}\sum_{t_0+1}^T (\boldsymbol{X}_t - \mathbf{B}_t((\mathbf{I}_m - \mathbf{I}_{t-1}^d)\boldsymbol{X}_{t-1} + \mathbf{I}_{t-1}^d(\mathbf{B}_{t-1}^*\boldsymbol{X}_{t_0} + \mathbf{f}_{t-1}^* + \mathbf{D}_{t-1}^*\boldsymbol{v})) - \mathbf{f}_{u,t} - \mathbf{D}_{u,t}\boldsymbol{v})^\top \mathbb{Q}_t \\
&(\boldsymbol{X}_t - \mathbf{B}_t((\mathbf{I}_m - \mathbf{I}_{t-1}^d)\boldsymbol{X}_{t-1} + \mathbf{I}_{t-1}^d(\mathbf{B}_{t-1}^*\boldsymbol{X}_{t_0} + \mathbf{f}_{t-1}^* + \mathbf{D}_{t-1}^* + \boldsymbol{v})) - \mathbf{f}_{u,t} - \mathbf{D}_{u,t}\boldsymbol{v}) \\
&- \frac{1}{2}\sum_{t_0}^T \log|\mathbf{Q}_t| - \frac{1}{2}(\boldsymbol{X}_{t_0} - \boldsymbol{\xi})^\top \mathbb{L}(\boldsymbol{X}_{t_0} - \boldsymbol{\xi}) - \frac{1}{2}\log|\boldsymbol{\Lambda}| - \frac{n}{2}\log 2\pi
\end{aligned}
\tag{195}
$$

where $\mathbf{B}^*$, $\mathbf{f}^*$ and $\mathbf{D}^*$ are defined in equation 190. $\mathbb{R}_t = \Xi_t^\top \mathbf{R}_t^{-1}\Xi_t$ and $\mathbb{Q}_t = \Phi_t^\top \mathbf{Q}_t^{-1}\Phi_t$, $\mathbb{L} = \Pi^\top\boldsymbol{\Lambda}^{-1}\Pi$. When $\boldsymbol{x}_{t_0}$ is treated as fixed, $\mathbb{L} = 0$ and the last line will drop out altogether, however in general some rows of $\boldsymbol{x}_{t_0}$ could be fixed and others stochastic.

We can see directly in equation 195 where $\boldsymbol{\upsilon}$ appears in the expected log-likelihood. Where $\mathbf{p}$ appears is less obvious because it depends on $\mathbf{F}$, which specifies which rows of $\boldsymbol{x}_{t_0}$ are fixed. From equation 192,

$$\mathrm{E}[\boldsymbol{X}_{t_0}] = (\mathbf{I}_m - \mathbf{I}_l^{(0)})\,\mathrm{E}[\boldsymbol{X}_{t_0}] + \mathbf{I}_l^{(0)}\boldsymbol{\xi}$$

and $\boldsymbol{\xi} = \mathbf{f}_\xi + \mathbf{D}_\xi \mathbf{p}$. Thus where $\mathbf{p}$ appears in the expected log-likelihood depends on the location of zero rows in $\mathbf{F}$ (and thus the zero rows in the indicator matrix $\mathbf{I}_l^{(0)}$). Recall that $\mathrm{E}[\boldsymbol{X}_{t_0}]$ appearing in the expected log-likelihood function is conditioned on the data so $\mathrm{E}[\boldsymbol{X}_{t_0}]$ in Ψ is not equal to $\boldsymbol{\xi}$ if $\boldsymbol{x}_{t_0}$ is stochastic.

The case where $\boldsymbol{x}_{t_0}$ is stochastic is a little odd because conditioned on $\boldsymbol{X}_{t_0} = \boldsymbol{x}_{t_0}$, $\boldsymbol{x}_t^d$ is deterministic even though $\boldsymbol{X}_0$ is a random variable in the model. Thus in the model, $\boldsymbol{x}_t^d$ is a random variable through $\boldsymbol{X}_{t_0}$. But when we do the partial differentiation step for the EM algorithm, we hold $\boldsymbol{X}$ at its expected value thus we are holding $\boldsymbol{X}_{t_0}$ at a specific value. We cannot do that and change $\mathbf{u}$ at the same time because once we fix $\boldsymbol{X}_{t_0}$ the $\boldsymbol{x}_t^d$ are deterministic functions of $\mathbf{u}$.

7.4 Logical constraints to ensure a consistent system of equations

We need to ensure that the model remains internally consistent when $\mathbf{R}$ or $\mathbf{Q}$ goes to zero and that we do not have an over- or under-constrained system.

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

$$\mathbf{H}_t\mathbf{R}_t = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & b & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{196}$$

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

$$\mathbf{Z}_{\text{bad}} = \begin{bmatrix} c & d & 0 \\ z(2,1) & z(2,2) & z(2,3) \\ z(3,1) & z(3,1) & z(3,1) \\ c & d & 0 \end{bmatrix}, \quad \mathbf{Z}_{\text{ok}} = \begin{bmatrix} c & 0 & 0 \\ z(2,1) & z(2,2) & z(2,3) \\ z(3,1) & z(3,1) & z(3,1) \\ c & d \neq 0 & 0 \end{bmatrix} \tag{197}$$

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

$$\begin{bmatrix} y_t(1) \\ y_t(4) \end{bmatrix} = \begin{bmatrix} cx_t(1) + dx_t(2) \\ cx_t(1) + dx_t(2) \end{bmatrix} \tag{198}$$

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

$$\begin{bmatrix} y_t(1) \\ y_t(4) \end{bmatrix} = \begin{bmatrix} cx_t(1) \\ cx_t(1) + dx_t(4) \end{bmatrix} \tag{199}$$

and that is solvable for any $y_t(1)$ and $y_t(4)$ combination. Notice that in the latter case, $x_t(1)$ and $x_t(2)$ are fully specified by $y_t(1)$ and $y_t(4)$.

7.4.1 Constraint 1: Z does not lead to an over-determined observation process

We need to ensure that a $\boldsymbol{x}_t$ exists for all $\boldsymbol{y}_t^{(0)}$ such that:

$$\mathrm{E}[\boldsymbol{Y}_t^{(0)}] = \mathbf{Z}^{(0)}\,\mathrm{E}[\boldsymbol{X}_t] + \mathbf{a}^{(0)}.$$

If $\mathbf{Z}^{(0)}$ is invertible, such a $\boldsymbol{x}_t$ certainly exists. But we do not require that only one $\boldsymbol{x}_t$ exists, simply that at least one exists. Thus the system can be under-constrained but not over-constrained. One way to test for this is to use the singular value decomposition (SVD) of $\mathbf{Z}^{(0)}$ ($\mathbf{Z}^{(0)}$ square). If the number of singular values of $\mathbf{Z}^{(0)}$ is less than the number of columns in $\mathbf{Z}$, which is the number of $\boldsymbol{x}$ rows, then $\mathbf{Z}^{(0)}$ specifies an over-constrained system ($y = Zx$[21]) Using the R language, you would test if the length of `svd(Z)$d` is less than than `dim(Z)[2]`. If $\mathbf{Z}^{(0)}$ specifies and under-determined system, some of the singular values would be equal

[21] This is the classic problem of solving the system of linear equations, which is standardly written $Ax = b$.

to 0 (within machine tolerance). It is possible that $\mathbf{Z}^{(0)}$ could specify both an over- and under-determined system at the same time. That is, the number of singular values could be less than the number of columns in $\mathbf{Z}^{(0)}$ and some of the singular values could be 0.

Doesn't a $\mathbf{Z}$ with more rows than columns automatically specify a over-determined system? No. Considered this $\mathbf{Z}$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{200}$$

This $\mathbf{Z}$ is fine, although obviously the last row of $\boldsymbol{y}$ will not hold any information about the $\boldsymbol{x}$. But it could have information about $\mathbf{R}$ and $\mathbf{a}$, which might be shared with the other $\boldsymbol{y}$, so we don't want to prevent the user from specifying a $\mathbf{Z}$ like this.

7.4.2 Constraint 2: the state processes are not over-constrained.

We also need to be concerned with the state process being over-constrained when both $\mathbf{Q} = 0$ and $\mathbf{R} = 0$ because we can have a situation where the constraint imposed by the observation process is at odds with the constraint imposed by the state process. Here is an example:

$$\begin{aligned} \boldsymbol{y}_t &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_t \\ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_t &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_{t-1} + \begin{bmatrix} w_1 \\ 0 \end{bmatrix}_{t-1} \end{aligned} \tag{201}$$

In this case, some of the x's are deterministic, $\mathbf{Q} = 0$ and not linked through $\mathbf{B}$ to a stochastic x, and the corresponding y are also deterministic. These cases will show up as errors in the Kalman filter/smoother because in the Kalman gain equation (equation 143e), the term $\mathbf{Z}_t \boldsymbol{V}_t^{t-1} \mathbf{Z}_t^\top$ will appear when $\mathbf{R} = 0$. We need to make sure that 0 rows in $\mathbf{B}_t$, $\mathbf{Z}_t$ and $\mathbf{Q}_t$ do not line up in such a way that 0 rows/cols do not appear in $\mathbf{Z}_t \boldsymbol{V}_t^{t-1} \mathbf{Z}_t^\top$ at the same place as 0 rows/cols in $\mathbf{R}$. In MARSS, this is checked by doing a pre-run of the Kalman smoother to see if it throws an error in the Kalman gain step.

8 EM algorithm modifications for degenerate models

The $\mathbf{R}$, $\mathbf{Q}$, $\mathbf{Z}$, and $\mathbf{a}$ update equations are largely unchanged. The real difficulties arise for the $\mathbf{u}$ and $\boldsymbol{\xi}$ update equations when $\mathbf{u}^{(0)}$ or $\boldsymbol{\xi}^{(0)}$ are estimated. For $\mathbf{B}$, I do not have a degenerate update equation, so I need to assume that $\mathbf{B}^{(0)}$ elements are fixed (not estimated).

8.1 R and Q update equations

The constrained update equations for $\mathbf{Q}$ and $\mathbf{R}$ work fine because their update equations do not involve any inverses of non-invertible matrices. However if $\mathbf{H}_t \mathbf{R}_t \mathbf{H}_t^\top$ is non-diagonal and there are missing values, then the $\mathbf{R}$ update equation involves $\widetilde{\mathbf{y}}_t$. That will involve the inverse of $\mathbf{H}_t \mathbf{R}_{11} \mathbf{H}_t^\top$ (section 6.2), which might have zeros on the diagonal. In that case, use the ∇_t modification that deals with such zeros (equation 150).

8.2 Z and a update equations

We need to deal with $\mathbf{Z}$ and $\mathbf{a}$ elements that appear in rows where the diagonal of $\mathbf{R} = 0$. These values will not appear in the likelihood function unless they also happen to also appear on the rows where the diagonal of $\mathbf{R}$ is not 0 (because they are constrained to be equal for example). However, in this case the $\mathbf{Z}^{(0)}$ and $\mathbf{a}^{(0)}$ are logically constrained by the equation

$$\boldsymbol{y}_t^{(0)} = \mathbf{Z}_t^{(0)} \, \mathrm{E}[\boldsymbol{x}_t] + \mathbf{a}_t^{(0)}.$$

Notice there is no $\boldsymbol{w}_t$ since $\mathbf{R} = 0$ for these rows. The $\mathrm{E}[\boldsymbol{x}_t]$ is ML estimate of $\boldsymbol{x}_t$ computed in the Kalman smoother from the parameter values at iteration i of the EM algorithm, so there is no information in this

equation for $\mathbf{Z}$ and $\mathbf{a}$ at iteration $i+1$. The nature of the smoother is that it will find the $\boldsymbol{x}_t$ that is most consistent with the data. For example if our $y = Zx + a$ equation looks like so

$$\begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} x, \tag{202}$$

there is no x that will solve this. However $x = 1$ is the closest (lowest squared error) and so this is the information in the data about x. The Kalman filter will use this and the relative value of $\mathbf{Q}$ and $\mathbf{R}$ to come up with the estimated x. In this case, $\mathbf{R} = 0$, so the information in the data will completely determine x and the smoother would return $x = 1$ regardless of the process equation.

The $\mathbf{a}$ and $\mathbf{Z}$ update equations require that $\sum_{t=1}^{T} \mathbf{D}_{t,a}^{\top} \mathbb{R}_t \mathbf{D}_{t,a}$ and $\sum_{t=1}^{T} \mathbf{D}_{t,z}^{\top} \mathbb{R}_t \mathbf{D}_{t,z}$ are invertible. If $\mathbf{Z}_t^{(0)}$ and $\mathbf{a}_t^{(0)}$ are fixed, this will be satisfied, however the restriction is a little less restrictive than that since it is possible that $\mathbb{R}_t$ does not have zeros on the diagonal in the same places so that the sum over t could be invertible while the individual values at t are not. The section on the summary of constraints has the test for this constraint.

The update equations also involve $\widetilde{\mathbf{y}}_t$, and the modified algorithm for $\widetilde{\mathbf{y}}_t$ when $\mathbf{H}_t$ has all zero rows will be needed. Other than that, the constrained update equations work (sections 5.2 and 5.7).

8.3 u update equation

Here I discuss the update for $\mathbf{u}$, or more specifically $\boldsymbol{v}$ which appears in $\mathbf{u}$, when $\mathbf{G}_t$ or $\mathbf{H}_t$ have zero rows. I require that $\mathbf{u}_t^{is}$ is not estimated. All the $\mathbf{u}_t^{is}$ are fixed values. The $\mathbf{u}_t^d$ may be estimated or more specifically there may be $\boldsymbol{v}$ in $\mathbf{u}_t^d$ that are estimated; $\mathbf{u}_t^d = \mathbf{f}_{u,t}^d + \mathbf{D}_{u,t}^d \boldsymbol{v}$.

For the constrained $\mathbf{u}$ update equation with deterministic $\boldsymbol{x}$'s takes the following form. It is similar to the unconstrained update equation except that that a part from the $\boldsymbol{y}$ part of the likelihood now appears:

$$\boldsymbol{v}_{j+1} = \left(\sum_{t=1}^{T} (\Delta_{t,2}^{\top} \mathbb{R}_t \Delta_{t,2} + \Delta_{t,4}^{\top} \mathbb{Q}_t \Delta_{t,4}) \right)^{-1} \times \left(\sum_{t=1}^{T} \left(\Delta_{t,2}^{\top} \mathbb{R}_t \Delta_{t,1} + \Delta_{t,4}^{\top} \mathbb{Q}_t \Delta_{t,3} \right) \right) \tag{203}$$

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

8.3.1 $\mathbf{u}^{(0)}$ is not estimated

When $\mathbf{u}^{(0)}$ is not estimated (since it is at some user defined value via $\mathbf{D}_u$ and $\mathbf{f}_u$), the part we are estimating, $\mathbf{u}^+$, only appears in the $\boldsymbol{x}$ part of the likelihood. The update equation for $\mathbf{u}$ remains equation 97.

8.3.2 $\mathbf{u}^d$ is estimated

The derivation of the update equation proceeds as usual. We need to take the partial derivative of Ψ^+ (equation 195) holding everything constant except $\boldsymbol{v}$, elements of which might appear in both $\mathbf{u}_t^d$ and $\mathbf{u}_t^s$ (but not $\mathbf{u}_t^{is}$ since I require that $\mathbf{u}_t^{is}$ has no estimated elements).

The expected log-likelihood takes the following form, where t_0 is the time where the initial state is defined ($t = 0$ or $t = 1$):

$$\begin{aligned} \Psi^+ = &-\frac{1}{2} \sum_{1}^{T} (\Delta_{t,1} - \Delta_{t,2}\boldsymbol{v})^{\top} \mathbb{R}_t (\Delta_{t,1} - \Delta_{t,2}\boldsymbol{v}) - \frac{1}{2} \sum_{1}^{T} \log |\mathbf{R}_t| \\ &-\frac{1}{2} \sum_{t_0+1}^{T} (\Delta_{t,3} - \Delta_{t,4}\boldsymbol{v})^{\top} \mathbb{Q}_t (\Delta_{t,3} - \Delta_{t,4}\boldsymbol{v}) - \frac{1}{2} \sum_{t_0+1}^{T} \log |\mathbf{Q}_t| \\ &-\frac{1}{2} (\boldsymbol{X}_{t_0} - \boldsymbol{\xi})^{\top} \mathbb{L} (\boldsymbol{X}_{t_0} - \boldsymbol{\xi}) - \frac{1}{2} \log |\boldsymbol{\Lambda}| - \frac{n}{2} \log 2\pi \end{aligned} \tag{204}$$

$\mathbb{L} = \mathbf{F}^\top \mathbf{\Lambda}^{-1}\mathbf{F}$. If $\boldsymbol{x}_{t_0}$ is treated as fixed, $\mathbf{F}$ is all zero and the line with $\mathbb{L}$ drops out. If some but not all $\boldsymbol{x}_{t_0}$ are treated as fixed, then only the stochastic rows appear in the last line. In any case, the last line does not contain $\boldsymbol{v}$, thus when we do the partial differentiation with respect to $\boldsymbol{v}$, this line drops out.

The Δ terms are defined as:

$$\begin{aligned}
\Delta_{t,1} &= \widetilde{\mathbf{y}}_t - \mathbf{Z}_t(\mathbf{I}_m - \mathbf{I}_t^d)\widetilde{\mathbf{x}}_t - \mathbf{Z}_t\mathbf{I}_t^d(\mathbf{B}_t^* \,\mathrm{E}[\boldsymbol{X}_{t_0}] + \mathbf{f}_t^*) - \mathbf{a}_t \\
\Delta_{t,2} &= \mathbf{Z}_t\mathbf{I}_t^d\mathbf{D}_t^* \\
\Delta_{t_0,3} &= 0_{m\times 1} \\
\Delta_{t,3} &= \widetilde{\mathbf{x}}_t - \mathbf{B}_t(\mathbf{I}_m - \mathbf{I}_{t-1}^d)\widetilde{\mathbf{x}}_{t-1} - \mathbf{B}_t\mathbf{I}_{t-1}^d(\mathbf{B}_{t-1}^* \,\mathrm{E}[\boldsymbol{X}_{t_0}] + \mathbf{f}_{t-1}^*) - \mathbf{f}_{t,u} \\
\Delta_{t_0,4} &= 0_{m\times m}\mathbf{D}_{1,u} \\
\Delta_{t,4} &= \mathbf{D}_{t,u} + \mathbf{B}_t\mathbf{I}_{t-1}^d\mathbf{D}_{t-1}^* \\
\mathrm{E}[\boldsymbol{X}_{t_0}] &= ((\mathbf{I}_m - \mathbf{I}_\lambda^{(0)})\widetilde{\boldsymbol{x}}_{t_0} + \mathbf{I}_\lambda^{(0)}\boldsymbol{\xi})
\end{aligned} \tag{205}$$

$\mathbf{I}_t^d$, $\mathbf{B}_t^*$, $\mathbf{f}_t*$, and $\mathbf{D}_t^*$ are defined in equation 190. The values of these at t_0 is special so that the math works out. The expectation (E) has been subsumed into the Δs since Δ_2 and Δ_4 do not involve $\boldsymbol{X}$ or $\boldsymbol{Y}$, so terms like $\boldsymbol{X}^\top\boldsymbol{X}$ never appear.

Take the derivative of this with respect to $\boldsymbol{v}$ and arrive at:

$$\boldsymbol{v}_{j+1} = \Big(\sum_{t=1}^{T}\Delta_{t,4}^\top\mathbb{Q}_t\Delta_{t,4} + \sum_{t=1}^{T}\Delta_{t,2}^\top\mathbb{R}_t\Delta_{t,2}\Big)^{-1} \times \left(\sum_{t=1}^{T}\Delta_{1,4}^\top\mathbb{Q}_t\Delta_{1,3} + \sum_{t=1}^{T}\Delta_{t,2}^\top\mathbb{R}_t\Delta_{t,1})\right) \tag{206}$$

8.4 $\boldsymbol{\xi}$ update equation

8.4.1 $\boldsymbol{\xi}$ is stochastic

This means that none of the rows of $\mathbf{F}$ (in $\mathbf{F}\lambda$) are zero, so $\mathbf{I}_\lambda^{(0)}$ is all zero and the update equation reduces to a constrained version of the classic $\boldsymbol{\xi}$ update equation:

$$\mathbf{p}_{j+1} = \big(\mathbf{D}_\xi^\top\mathbf{\Lambda}^{-1}\mathbf{D}_\xi\big)^{-1}\mathbf{D}_\xi^\top\mathbf{\Lambda}^{-1}(\,\mathrm{E}[\boldsymbol{X}_{t_0}] - \mathbf{f}_\xi) \tag{207}$$

8.4.2 $\boldsymbol{\xi}^{(0)}$ is not estimated

When $\boldsymbol{\xi}^{(0)}$ is not estimated (because you fixed it as some value), we do not need to take the partial derivative with respect to $\boldsymbol{\xi}^{(0)}$ since we will not be estimating it. Thus the update equation is unchanged from the constrained update equation.

8.4.3 $\boldsymbol{\xi}^{(0)}$ is estimated

Using the same approach as for $\mathbf{u}$ update equation, we take the derivative of 195 with respect to $\mathbf{p}$ where $\boldsymbol{\xi} = \mathbf{f}_\xi + \mathbf{D}_\xi\mathbf{p}$. Ψ^+ will take the following form:

$$\begin{aligned}
\Psi^+ = &\\
&-\frac{1}{2}\sum_{t=1}^{T}(\Delta_{t,5} - \Delta_{t,6}\mathbf{p})^\top\mathbb{R}_t(\Delta_{t,5} - \Delta_{t,6}\mathbf{p}) - \frac{1}{2}\sum_{1}^{T}\log|\mathbf{R}_t| \\
&-\frac{1}{2}\sum_{t=1}^{T}(\Delta_{t,7} - \Delta_{t,8}\mathbf{p})^\top\mathbb{Q}_t(\Delta_{t,7} - \Delta_{t,8}\mathbf{p}) - \frac{1}{2}\sum_{1}^{T}\log|\mathbf{Q}_t| \\
&-\frac{1}{2}(\,\mathrm{E}[\boldsymbol{X}_{t_0}] - \mathbf{f}_\xi - \mathbf{D}_\xi\mathbf{p})^\top\mathbb{L}(\,\mathrm{E}[\boldsymbol{X}_{t_0}] - \mathbf{f}_\xi - \mathbf{D}_\xi\mathbf{p}) - \frac{1}{2}\log|\mathbf{\Lambda}| \\
&-\frac{n}{2}\log 2\pi
\end{aligned} \tag{208}$$

The Δ's are defined as follows using $\mathrm{E}[\boldsymbol{X}_{t_0}] = (\mathbf{I}_m - \mathbf{I}_l^{(0)})\widetilde{\boldsymbol{x}}_{t_0} + \mathbf{I}_l^{(0)}\boldsymbol{\xi}$ where it appears in $\mathbf{I}_t^d \mathrm{E}[\boldsymbol{X}_t]$.

$$\begin{aligned}
\Delta_{t,5} &= \widetilde{\mathbf{y}}_t - \mathbf{Z}_t(\mathbf{I}_m - \mathbf{I}_t^d)\widetilde{\mathbf{x}}_t - \mathbf{Z}_t\mathbf{I}_t^d(\mathbf{B}_t^*((\mathbf{I}_m - \mathbf{I}_\lambda^{(0)})\widetilde{\boldsymbol{x}}_{t_0} + \mathbf{I}_\lambda^{(0)}\mathbf{f}_\xi) + \mathbf{u}_t^*) - \mathbf{a}_t \\
\Delta_{t,6} &= \mathbf{Z}_t\mathbf{I}_l^d\mathbf{B}_t^*\mathbf{I}_\lambda^{(0)}\mathbf{D}_\xi \\
\Delta_{t_0,7} &= 0_{m\times 1} \\
\Delta_{t,7} &= \widetilde{\mathbf{x}}_t - \mathbf{B}_t(\mathbf{I}_m - \mathbf{I}_{t-1}^d)\widetilde{\mathbf{x}}_{t-1} - \mathbf{B}_t\mathbf{I}_{t-1}^d(\mathbf{B}_{t-1}^*((\mathbf{I}_m - \mathbf{I}_l^{(0)})\widetilde{\boldsymbol{x}}_{t_0} + \mathbf{I}_\lambda^{(0)}\mathbf{f}_\xi) + \mathbf{u}_{t-1}^*) - \mathbf{u}_t \\
\Delta_{t_0,8} &= 0_{m\times m}\mathbf{D}_\xi \\
\Delta_{t,8} &= \mathbf{B}_t\mathbf{I}_{t-1}^d\mathbf{B}_{t-1}^*\mathbf{I}_\lambda^{(0)}\mathbf{D}_\xi \\
\mathbf{u}_t^* &= \mathbf{f}_t^* + \mathbf{D}_t^*\boldsymbol{v}
\end{aligned} \tag{209}$$

The expectation can be pulled inside the Δs since the Δs in front of $\mathbf{p}$ do not involve $\boldsymbol{X}$ or $\boldsymbol{Y}$.

Take the derivative of this with respect to $\mathbf{p}$ and arrive at:

$$\begin{aligned}
\mathbf{p}_{j+1} = \Big(\sum_{t=1}^T \Delta_{t,8}^\top\mathbb{Q}_t\Delta_{t,8} + \sum_{t=1}^T \Delta_{t,6}^\top\mathbb{R}_t\Delta_{t,6} + \mathbf{D}_\xi^\top\mathbb{L}\mathbf{D}_\xi\Big)^{-1}\times \\
\Big(\sum_{t=1}^T \Delta_{1,8}^\top\mathbb{Q}_t\Delta_{1,7} + \sum_{t=1}^T \Delta_{t,6}^\top\mathbb{R}_t\Delta_{t,5} + \mathbf{D}_\xi^\top\mathbb{L}(\mathrm{E}[\boldsymbol{X}_{t_0}] - \mathbf{f}_\xi)\Big)
\end{aligned} \tag{210}$$

8.4.4 When $\mathbf{H}_t$ has 0 rows in addition to $\mathbf{G}_t$

When $\mathbf{H}_t$ has all zero rows, some of the $\mathbf{p}$ or $\boldsymbol{v}$ may constrained by the model, but these constraints do not appear in Ψ^+ since $\mathbb{R}_t$ zeros out those constraints. For example, if H_t is all zeros and $\boldsymbol{x}_1 \equiv \boldsymbol{\xi}$, then $\boldsymbol{\xi}$ is constrained to equal $\mathbf{Z}^{-1}(\widetilde{\mathbf{y}}_1 - \mathbf{a}_1)$.

The model needs to be internally consistent and we need to be able to estimate all the $\mathbf{p}$ and the $\boldsymbol{v}$. Rather than try to estimate the correct $\mathbf{p}$ and $\boldsymbol{v}$ to ensure internal consistency of the model with the data when some of the $\mathbf{H}_t$ have 0 rows, I test by running the Kalman filter with the degenerate variance modification (in particular the modification for $\mathbf{F}$ with zero rows is critical) before starting the EM algorithm. Then I test that $\widetilde{\mathbf{y}}_t - \mathbf{Z}_t\widetilde{\mathbf{x}}_t - \mathbf{a}_t$ is all zeros. If it is not, within machine accuracy, then there is a problem. This is reported and the algorithm stopped[22]

I also test that $\big(\sum_{t=1}^T \Delta_{t,8}^\top\mathbb{Q}_t\Delta_{t,8} + \sum_{t=1}^T \Delta_{t,6}^\top\mathbb{R}_t\Delta_{t,6} + \mathbf{D}_\xi^\top\mathbb{L}\mathbf{D}_\xi\big)$ is invertible to ensure that all the $\mathbf{p}$ can be solved for, and I test that $\big(\sum_{t=1}^T \Delta_{t,4}^\top\mathbb{Q}_t\Delta_{t,4} + \sum_{t=1}^T \Delta_{t,2}^\top\mathbb{R}_t\Delta_{t,2}\big)$ is invertible so that all the $\boldsymbol{v}$ can be solved for. If errors are present, they should be apparent in iteration 1, are reported and the EM algorithm stopped.

8.5 $\mathbf{B}^{(0)}$ update equation for degenerate models

I do not have an update equation for $\mathbf{B}^{(0)}$ and for now, I side-step this problem by requiring that any $\mathbf{B}^{(0)}$ terms are fixed.

9 Kalman filter and smoother modifications for degenerate models

9.1 Modifications due to degenerate R and Q

[1/1/2012 note. These modifications mainly have to do with inverses that appear in the Shumway and Stoffer's presentation of the Kalman filter. The MARSS package uses Koopman's smoother algorithm which avoids these inverses altogether however these appear in the MARSSkfss() function (the Shumway and Stoffer implementation).]

In principle, when either $\mathbf{G}_t\mathbf{Q}_t$ or $\mathbf{H}_t\mathbf{R}_t$ has zero rows, the standard Kalman filter/smoother equations would still work and provide the correct state outputs and likelihood. In practice however errors will be

[22] In some cases, it is easy to determine the correct $\boldsymbol{\xi}$. For example, when $\mathbf{H}_t$ is all zero rows, $t_0 = 1$ and there is no missing data at time $t = 1$, $\boldsymbol{\xi} = \mathbf{Z}^*(\boldsymbol{y}_1 - \mathbf{a}_1)$, where $\mathbf{Z}^*$ is the pseudoinverse. One would want to use the SVD pseudoinverse calculation in case $\mathbf{Z}$ leads to an under-constrained system (some of the singular values of $\mathbf{Z}$ are 0).

generated because under certain situations, one of the matrix inverses in the Kalman filter/smoother equations will involve a matrix with a zero on the diagonal and this will lead to the computer code throwing an error.

When $\mathbf{H}_t\mathbf{R}_t$ has zero rows, problems arise in the Kalman update part of the Kalman filter. The Kalman gain is

$$\mathbf{K}_t = \boldsymbol{V}_t^{t-1}(\mathbf{Z}_t^*)^\top(\mathbf{Z}_t^*\boldsymbol{V}_t^{t-1}(\mathbf{Z}_t^*)^\top + \mathbf{H}_t\mathbf{R}_t^*\mathbf{H}_t^\top)^{-1} \tag{211}$$

Here, $\mathbf{Z}_t^*$ is the missing values modified $\mathbf{Z}_t$ matrix with the i-th rows zero-ed out if the i-th element of $\boldsymbol{y}_t$ is missing (section 6.1, equation 145). Thus if the i-th element of $\boldsymbol{y}_t$ is missing and the i-th row of $\mathbf{H}_t$ is zero, the (i,i) element of $(\mathbf{Z}_t^*\boldsymbol{V}_t^{t-1}(\mathbf{Z}_t^*)^\top + \mathbf{H}_t\mathbf{R}_t^*\mathbf{H}_t^\top)$ 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 $\boldsymbol{V}_1^0$ will be a $m \times m$ matrix of zeros. Again in this situation $(\mathbf{Z}_t^*\boldsymbol{V}_t^{t-1}(\mathbf{Z}_t^*)^\top + \mathbf{H}_t\mathbf{R}_t^*\mathbf{H}_t^\top)$ will have zeros at any (i,i) elements where the i-th row of $\mathbf{H}_t$ 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 $\left(\mathbf{Z}_t^*\boldsymbol{V}_t^{t-1}(\mathbf{Z}_t^*)^\top + \mathbf{H}_t\mathbf{R}_t^*\mathbf{H}_t^\top\right)^{-1}$ in equation 211 with

$$(\boldsymbol{\Omega}_t^{(1)})^\top\left(\boldsymbol{\Omega}_t^{(1)}(\mathbf{Z}_t^*\boldsymbol{V}_t^{t-1}(\mathbf{Z}_t^*)^\top + \mathbf{H}_t\mathbf{R}_t^*\mathbf{H}_t^\top)(\boldsymbol{\Omega}_t^{(1)})^\top\right)^{-1}\boldsymbol{\Omega}_t^{(1)} \tag{212}$$

Wrapping in $\boldsymbol{\Omega}_t^{(1)}(\boldsymbol{\Omega}_t^{(1)})^\top$ gets rid of all the zero rows/columns in $\mathbf{Z}_t'\boldsymbol{V}_t^{t-1}(\mathbf{Z}_t')^\top + \mathbf{H}_t\mathbf{R}_t'\mathbf{H}_t^\top$, 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 $\boldsymbol{V}_1^0 = \mathbf{0}$, set $\mathbf{K}_1 = \mathbf{0}$ without computing the inverse (see equation 211 where $\boldsymbol{V}_1^0$ appears on the left).

There is also a numerical issue to deal with. When the i-th row of $\mathbf{H}_t$ is 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 $\boldsymbol{V}_t^t$ must be zero because given $y_t(i)$, $x_t(k)$ is known (is fixed) and its variance, $\boldsymbol{V}_t^t(k,k)$, is zero. Because $\mathbf{K}_t$ is computed using a numerical estimate of the inverse, the standard $\boldsymbol{V}_t^t$ 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 $\boldsymbol{V}_t^t$ is computed using the normal Kalman update equation, we will want to explicitly zero out the k rows and columns in the filter.

When $\mathbf{G}_t$ has zero rows, then we might also have similar numerical errors in $\mathbf{J}$ in the Kalman smoother. The $\mathbf{J}$ equation is

$$\begin{aligned}\mathbf{J}_t &= \boldsymbol{V}_{t-1}^{t-1}\mathbf{B}_t^\top(\boldsymbol{V}_t^{t-1})^{-1} \\ &\text{where } \boldsymbol{V}_t^{t-1} = \mathbf{B}_t\boldsymbol{V}_{t-1}^{t-1}\mathbf{B}_t^\top + \mathbf{G}_t\mathbf{Q}_t\mathbf{G}_t^\top\end{aligned} \tag{213}$$

If there are zeros on the diagonals of ($\boldsymbol{\Lambda}$ and/or $\mathbf{B}_t$) and zero rows in $\mathbf{G}_t$ and these zeros line up, then if the $\mathbf{B}_t^{(0)}$ and $\mathbf{B}_T^{(1)}$ elements in $\mathbf{B}_t$ are blocks[23], there will be zeros on the diagonal of $\boldsymbol{V}_t^t$. Thus there will be zeros on the diagonal of $\boldsymbol{V}_t^{t-1}$ and it cannot be inverted. In this case, the corresponding elements of $\boldsymbol{V}_t^T$ need to be zero since what's happening is that those elements are deterministic and thus have 0 variance.

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

$$(\boldsymbol{\Omega}_{Vt}^+)^\top\left(\boldsymbol{\Omega}_{Vt}^+(\boldsymbol{V}_t^{t-1})(\boldsymbol{\Omega}_{Vt}^+)^\top\right)^{-1}\boldsymbol{\Omega}_{Vt}^+ \tag{214}$$

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

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

9.2 Modifications due to fixed initial states

When the initial state of $\boldsymbol{x}$ is fixed, then it is a bit like $\boldsymbol{\Lambda} = 0$ although actually $\boldsymbol{\Lambda}$ does not appear in the model and $\boldsymbol{\xi}$ has a different interpretation.

When the initial state of $\boldsymbol{x}$ is treated as stochastic, then if $t_0 = 0$, $\boldsymbol{\xi}$ is the expected value of $\boldsymbol{x}_0$ conditioned on no data. In the Kalman filter this means $\boldsymbol{x}_0^0 = \boldsymbol{\xi}$ and $\boldsymbol{V}_0^0 = \boldsymbol{\Lambda}$; in words, the expected value of $\boldsymbol{x}_0$ conditioned on $\boldsymbol{y}_0$ is $\boldsymbol{\xi}$ and the variance of $\boldsymbol{x}_0^0$ conditioned on $\boldsymbol{y}_0$ is $\boldsymbol{\Lambda}$. When $t_0 = 1$, then $\boldsymbol{\xi}$ is the expected value of $\boldsymbol{x}_1$ conditioned on no data. In the Kalman filter this means $\boldsymbol{x}_1^0 = \boldsymbol{\xi}$ and $\boldsymbol{V}_1^0 = \boldsymbol{\Lambda}$. Thus where $\boldsymbol{\xi}$ and $\boldsymbol{\Lambda}$ appear in the Kalman filter equations is different depending on t_0; the $\boldsymbol{x}_t^t$ and $\boldsymbol{V}_t^t$ initial condition versus the $\boldsymbol{x}_t^{t-1}$ and $\boldsymbol{V}_t^{t-1}$ initial condition.

When some or all of the $\boldsymbol{x}_{t_0}$ are fixed, denoted the $\mathbf{I}_\lambda^{(0)}\boldsymbol{x}_{t_0}$, the fixed values are not a random variables. While technically speaking, the expected value of a fixed value does not exist, we can think of it as a random variable with a probability density function with all the weight on the fixed value. Thus $\mathbf{I}_\lambda^{(0)}\,\mathrm{E}[\boldsymbol{x}_{t_0}] = \boldsymbol{\xi}$ regardless of the data. The data have no information for $\mathbf{I}_\lambda^{(0)}\boldsymbol{x}_{t_0}$ since we fix $\mathbf{I}_\lambda^{(0)}\boldsymbol{x}_{t_0}$ at $\mathbf{I}_\lambda^{(0)}\boldsymbol{\xi}$. If $t_0 = 0$, we initialize the Kalman filter as usual with $\boldsymbol{x}_0^0 = \boldsymbol{\xi}$ and $\boldsymbol{V}_0^0 = \mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top$, where the fixed $\boldsymbol{x}_{t_0}$ rows correspond to the zero row/columns in $\mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top$. The Kalman filter will return the correct expectations even when some of the diagonals of $\mathbf{H}\mathbf{R}\mathbf{H}^\top$ or $\mathbf{G}\mathbf{Q}\mathbf{G}^\top$ are 0—with the constraint that we have no purely deterministic elements in the model (meaning there are no errors terms from either $\mathbf{R}$ or $\mathbf{Q}$).

When $t_0 = 1$, $\mathbf{I}_\lambda^{(0)}\boldsymbol{x}_1^0$ and $\mathbf{I}_l^{(0)}\boldsymbol{x}_1^1 = \boldsymbol{\xi}$ regardless of the data and $\boldsymbol{V}_1^0 = \mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top$ and $\boldsymbol{V}_1^1 = \mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top$, where the fixed rows of $\boldsymbol{x}_1$ correspond with the 0 row/columns in $\mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top$. We also set $\mathbf{I}_\lambda^{(0)}\mathbf{K}_1$, meaning the rows of $\boldsymbol{x}_1$ that are fixed, to all zero because $\mathbf{K}_1$ is the information in $\boldsymbol{y}_1$ regarding $\boldsymbol{x}_1$ and there is no information in the data regarding the values of $\boldsymbol{x}_1$ that are fixed to equal $\mathbf{I}_\lambda^{(0)}\boldsymbol{\xi}$.

With $\boldsymbol{V}_1^1$, $\boldsymbol{x}_1^1$ and $\mathbf{K}_1$ set to their correct initial values, the normal Kalman filter equations will work fine. However it is possible for the data at $t = 1$ to be inconsistent with the model if the rows of $\boldsymbol{y}_1$ corresponding to any zero row/columns in $\mathbf{Z}_1\mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top\mathbf{Z}_1^\top + \mathbf{H}_1\mathbf{R}_1\mathbf{H}_1^\top$ are not equal to $\mathbf{Z}_1\boldsymbol{\xi} + \mathbf{a}_1$. Here is a trivial example, let the model be $x_t = x_{t-1} + w_t$, $y_t = x_t$, $x_1 = 1$. Then if y_1 is anything except 1, the model is impossible. Technically, the likelihood of x_1 conditioned on $Y_1 = y_1$ does not exist since neither x_1 nor y_1 are realizations of a random variable (since they are fixed), so when the likelihood is computed using the innovations form of the likelihood, the $t = 1$ does not appear, at least for those $\boldsymbol{y}_1$ corresponding to any zero row/columns in $\mathbf{Z}_1\mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^\top\mathbf{Z}_1^\top + \mathbf{H}_1\mathbf{R}_1\mathbf{H}_1^\top$. Thus these internal inconsistencies would neither provoke an error nor cause Inf to be returned for the likelihood. In the MARSS package, the Kalman filter has been modified to return LL=Inf and an error.

10 Summary of requirements for degenerate models

Below are discussed the update equations for the different parameters. Here I summarize the constraints that are scattered throughout these subsections. These requirements are coded into the function MARSSkemcheck() in the MARSS package but some tests must be repeated in the function degen.test(), which tests if any of the $\mathbf{R}$ or $\mathbf{Q}$ diagonals can be set to zero if it appears they are going to zero. A model that is allowed when $\mathbf{R}$ and $\mathbf{Q}$ are non-zero, might be disallowed if $\mathbf{R}$ or $\mathbf{Q}$ diagonals were to be set to zero. degen.test() does this check.

- $(\mathbf{I}_m \otimes \mathbf{I}_r^{(0)}\mathbf{Z}_t\mathbf{I}_q^{(0)})\mathbf{D}_{t,b}$, is all zeros. If there is a all zero row in $\mathbf{H}_t$ and it is linked (through $\mathbf{Z}$) to a all zero row in $\mathbf{G}_t$, then the corresponding $\mathbf{B}_t$ elements are fixed instead of estimated. Corresponding $\mathbf{B}$ rows means those rows in $\mathbf{B}$ where there is a non-zero column in $\mathbf{Z}$. We need $\mathbf{I}_r^{(0)}\mathbf{Z}_t\mathbf{I}_q^{(0)}\mathbf{B}_t$ to only specify fixed $\mathbf{B}_t$ elements, which means $\mathrm{vec}(\mathbf{I}_r^{(0)}\mathbf{Z}_t\mathbf{I}_q^{(0)}\mathbf{B}_t\mathbf{I}_m)$ only specifies fixed values. This in turn leads to the condition above. MARSSkemcheck()
- $(\mathbf{I}_1 \otimes \mathbf{I}_r^{(0)}\mathbf{Z}_t\mathbf{I}_q^{(0)})\mathbf{D}_{t,u}$ is all zeros; if there is a all zero row in $\mathbf{H}_t$ and it is linked (through $\mathbf{Z}_t$) to a all zero row in $\mathbf{G}_t$, then the corresponding $\mathbf{u}_t$ elements are fixed instead of estimated. MARSSkemcheck()
- $(\mathbf{I}_m \otimes \mathbf{I}_r^{(0)})\mathbf{D}_{t,z}$, where is all zeros; if y has no observation error, then the corresponding $\mathbf{Z}_t$ rows are fixed values. $(\mathbf{I}_m \otimes \mathbf{I}_r^{(0)})$ is a diagonal matrix with 1s for the rows of $\mathbf{D}_{t,z}$ that correspond to elements of $\mathbf{Z}_t$ on the $R = 0$ rows. MARSSkemcheck()

- $(\mathbf{I}_1 \otimes \mathbf{I}_r^{(0)})\mathbf{D}_{t,a}$ is all zeros; if y has no observation error, then the corresponding $\mathbf{a}_t$ rows are fixed values. MARSSkemcheck()
- $(\mathbf{I}_m \otimes \mathbf{I}_q^{(0)})\mathbf{D}_{t,b}$ is all zeros. This means $\mathbf{B}^{(0)}$ (the whole row) is fixed. While $\mathbf{B}^d$ could potentially be estimated potentially, my derivation assumes it is not. MARSSkemcheck()
- $(\mathbf{I}_1 \otimes \mathbf{I}_{q,t>m}^{is})\mathbf{D}_{t,u}$ is all zeros. This means $\mathbf{u}^{is}$ is fixed. Here is is defined as those rows that are indirectly stochastic at time m, where m is the dimension of $\mathbf{B}$; it can take up to m steps for the is rows to be connected to the s rows through $\mathbf{B}$. MARSSkemcheck()
- If $\mathbf{u}^{(0)}$ or $\boldsymbol{\xi}^{(0)}$ are being estimated, then the adjacency matrices defined by $\mathbf{B}_t \neq 0$ are not time-varying. This means that the locations of the 0s in $\mathbf{B}_t$ are not changing over time. $\mathbf{B}_t$ however may be time-varying. MARSSkemcheck()
- $\mathbf{I}_q^{(0)}$ and $\mathbf{I}_r^{(0)}$ are time invariant (an imposed assumption). This means that the location of the 0 rows in $\mathbf{G}_t$ and $\mathbf{H}_t$ (and thus in $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$) are not changing through time. It would be easy enough to allow $\mathbf{I}_r^{(0)}$ to be time varying, but to make my derivation easier, I assume it is time constant.
- $\mathbf{Z}_t^{(0)}$ in $\mathrm{E}[\boldsymbol{Y}_t^{(0)}] = \mathbf{Z}_t^{(0)}\,\mathrm{E}[\boldsymbol{X}_t] + \mathbf{a}_t^{(0)}$ does not imply an over-determined system of equations. Because the $\boldsymbol{v}_t$ rows are zero for the (0) rows of $\boldsymbol{y}$, it must be possible for this equality to hold. This means that $\mathbf{Z}_t^{(0)}$ cannot specify an over-determined system although an underdetermined system is ok. The check is in MARSSkfss() since the fully-specified x need to be known for the MARSSkfss() filter. If $\mathbf{Z}_t^{(0)}$ is square, its inverse is attempted and if that throws and error an error is reported (re over-constrained model). The function to find the fully determined $\boldsymbol{x}$ is fully.det.x() in the utility functions.
- The state process cannot be over-determined via constraints imposed from the deterministic observation process ($\mathbf{R} = 0$) and the deterministic state process ($\mathbf{Q} = 0$). If this is the case the Kalman gain equation (in the Kalman filter) will throw an error. Checked in MARSS() via call to MARSSkf() before fitting call; degen.test(), in MARSSkem() will also test via MARSSkf call if some R or Q are attempted to be set to 0. If B or Z changes during kem or optim iterations such that this constraint does not hold, then algorithm will exit with an error message.
- The location of the 0s in $\mathbf{B}$ are time-invariant. The $\mathbf{B}$ can be time-varying but not the location of 0s. Also, I want $\mathbf{B}$ to be such that once a row becomes indirectly stochastic is stays that way. For example, if $\mathbf{B} = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, then row 2 flips back and forth from being indirectly stochastic to deterministic.

The dimension of the identity matrices in the above constraints is given by the subscript on $\mathbf{I}$ except when it is implicit.

11 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. However an initial conditions search should be done before reporting final estimates for an analysis. In our papers on the distributional properties of MARSS parameter estimates, we rarely found that an initial conditions search changed the estimates—except in cases where $\mathbf{Z}$ and $\mathbf{B}$ are estimated as unconstrained and as the fraction of missing data in the data set became large.

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 $\boldsymbol{\Lambda}$ 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 $\boldsymbol{x}_0$ (or $\boldsymbol{x}_1$) and is then estimated as a parameter; $\boldsymbol{\Lambda}$ is held fixed at 0. In case 2, $\boldsymbol{\xi}$ and $\boldsymbol{\Lambda}$ specify

the mean and variance of $\boldsymbol{X}_0$ (or $\boldsymbol{X}_1$) respectively. Neither are estimated; instead, they are specified as part of the model.

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

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

- Treat $\boldsymbol{x}_0$ as an estimated parameter and set $\boldsymbol{V}_0$=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 $\boldsymbol{V}_0$ you use is telling the algorithm what correlation structure to use. If there is a mismatch between the correlation structure in the prior and the correlation structure implied by the MLE model, you will not be escaping the prior problem. But sometimes you will know your implied correlation structure. For example, you may know that the $\boldsymbol{x}$'s are independent or you may be able to solve for the stationary distribution a priori if your stationary distribution is not a function of the parameters you are trying to estimate. Other times you are estimating a parameter that determines the correlation structure (like $\mathbf{B}$) and you will not know a priori what the correlation structure is.

In some cases, the update equation for one parameter needs other parameters. Technically, the Kalman filter/smoother should be run between each parameter update, however following Ghahramani and Hinton (1996) the default MARSS algorithm skips this step (unless the user sets `control$safe=TRUE`) and each updated parameter is used for subsequent update equations. If you see warnings that the log-likelihood drops, then try setting `control$safe=TRUE`. This will increase computation time greatly.

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

References

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

Borman, S. (2009). The expectation maximization algorithm - a short tutorial.

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

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

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

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

Koopman, S. and Ooms, M. (2011). *Forecasting economic time series using unobserved components time series models*, pages 129–162. Oxford University Press, Oxford.

Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika*, 80(1):117–126.

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.