Computation of Standardized Residuals for MARSS Models

E. E. Holmes*

original 31 Oct 2014
added normalization calculations, 20 Oct 2020

Abstract

This report walks through a derivation of the variance-covariance matrix for the joint conditional model and state residuals for multivariate autoregressive Gaussian state-space (MARSS) models. The bulk of the report focuses on 'smoothations' residuals (Harvey et al., 1998), which are the residuals conditioned on all the data $t = 1$ to T. The final part of the report covers 'innovations' residuals, which are residuals conditioned on the data $t = 1$ to $t - 1$.

The MARSS model can be written: $\boldsymbol{x}_t = \mathbf{B}\boldsymbol{x}_{t-1} + \mathbf{u} + \boldsymbol{w}_t$, $\boldsymbol{y}_t = \mathbf{Z}\boldsymbol{x}_t + \mathbf{z} + \boldsymbol{v}_t$, where $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are independent multivariate Gaussian error-terms with variance-covariance matrices $\mathbf{Q}_t$ and $\mathbf{R}_t$ respectively. The joint conditional residuals are the $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ conditioned on the observed data, which may be incomplete (missing values). Harvey, Koopman and Penzer (1998) show a recursive algorithm for the smoothation residuals (residuals conditioned on all the data). I show an alternate algorithm to compute these residuals using the conditional variances of the states and the conditional covariance between unobserved data and states. This allows one to compute the variance of un-observed smoothation residuals (residuals associated with missing or left-out data), which is needed for leave-one-out cross-validation tests using smoothation residuals. I show how to modify the Harvey et al. algorithm in the case of missing values and how to modify it to return the non-normalized conditional residuals.

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

citation: Holmes, E. E. 2014. Computation of standardized residuals for (MARSS) models. Technical Report. arXiv:1411.0045

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

1 Overview of MARSS residuals

This report discusses the computation of the variance of the conditional model and state residuals for MARSS models of the form:

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

The state and model residuals are respectively

$$\begin{aligned} \boldsymbol{w}_t &= \boldsymbol{x}_t - \mathbf{B}_t\boldsymbol{x}_{t-1} - \mathbf{u}_t \\ \boldsymbol{v}_t &= \boldsymbol{y}_t - \mathbf{Z}_t\boldsymbol{x}_t - \mathbf{a}_t. \end{aligned} \tag{2}$$

The model (and state) residuals are a random variables since $\boldsymbol{y}_t$ and $\boldsymbol{x}_t$ are drawn from the joint multivariate distribution of $\boldsymbol{Y}_t$ and $\boldsymbol{X}_t$ defined by the MARSS equations (Equation 1). The unconditional[1] variance of the model residuals is

$$\text{var}_{XY}[\boldsymbol{V}_t] = \text{var}_{XY}[\boldsymbol{Y}_t - (\mathbf{Z}_t\boldsymbol{X}_t + \mathbf{a}_t)] = \mathbf{R}_t \tag{3}$$

based on the distribution of $\boldsymbol{V}_t$ in Equation 1. var_{XY} indicates that the integration is over the joint unconditional distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Once we have data, $\mathbf{R}_t$ is not the variance-covariance matrix of our model residuals because our residuals are now conditioned[2] on a set of observed data. There are two types of conditional model residuals used in MARSS analyses: innovations and smoothations. Innovations are the model residuals at time t using the expected value of $\boldsymbol{X}_t$ conditioned on the data from 1 to $t-1$. Smoothations are the model residuals using the expected value of $\boldsymbol{X}_t$ conditioned on all the data, $t = 1$ to T. Smoothations are used in computing standardized residuals for outlier and structural break detection (Harvey and Koopman, 1992; Harvey et al., 1998; de Jong and Penzer, 1998; Commandeur and Koopman, 2007).

It should be noted that all the calculations discussed here are conditioned on the MARSS parameters: **B**, **Q**, **U**, **R**, **Z** and **A**. These are treated as known. This is different than standard discussions of residual distributions for linear models where the uncertainty in the model parameters enters into the calculations (as it enters into the calculation of the influence of $\boldsymbol{y}$ on the expected (or fitted) value of $\boldsymbol{y}$). In the calculations in this report, $\boldsymbol{y}$ does not affect the estimates of the parameters (which are fixed, perhaps at estimated values) but does affect the expected value of $\boldsymbol{Y}_t$ by affecting the estimate of the expected value and variance of $\boldsymbol{X}_t$.

2 Distribution of MARSS smoothation residuals

This section discusses computation of the variance of the model and state residuals conditioned on all the data from $t = 1$ to T. These MARSS residuals are often used for outlier detection and shock detection, and in this case you only need the distribution of the model residuals for the observed values. However if you wanted to do a leave-one-out cross-validation, you would need to know the distribution of the residuals for data points you left out (treated as unobserved). The equations in this report give you the former and the latter, while the algorithm by Harvey et al. (1998) gives only the former. These equations for residuals for 'left-out' data are different that other (typical) discussions of state-space cross-validation (de Jong, 1988) in that they are conditioned on all the data (smoothations residuals) rather than conditioned on data up to $t-1$ (innovations residuals).

2.1 Notation and relations

Throughout, I follow the convention that capital letters are random variables and small letters are a realization from the random variable. This only applies to random variables; parameters are not random variables[3].

[1] meaning not conditioning on any particular set of observed data but rather taking the expectation across all possible values of $\boldsymbol{y}_t$ and $\boldsymbol{x}_t$.

[2] 'conditioned' means that the probability distribution of the residual has changed. The distribution is now the distribution given that $\boldsymbol{Y} = \boldsymbol{y}$, say. Expectations and variances var[] are integrals over the value that a random variable might take multiplied by the probability of that value. When presenting an 'expectation', the probability distribution is normally implicit but for derivations involving conditional expectations, it is important to be explicit about the distribution that is being integrated over.

[3] in a frequentist framework

Parameters are shown in Roman font while while random variables are bold slanted font. Parameters written as capital letters are matrices, while parameters written in small letters are strictly column matrices.

In this report, the distribution over which the integration is done in an expectation or variance is given by the subscript; e.g., $\mathrm{E}_A[f(A)]$ indicates an unconditional expectation over the distribution of A without conditioning on another random variable while $\mathrm{E}_{A|b}[f(A)|b]$ would indicate an expectation over the distribution of A conditioned on $B = b$; presumably A and B are not independent otherwise $B = b$ would have no effect on A. $\mathrm{E}_{A|b}[f(A)|b]$ is a fixed value, not random. It is the expected value when $B = b$. In contrast, $\mathrm{E}_{A|B}[f(A)|B]$ denotes the random variable over all the possible $\mathrm{E}_{A|b}[f(A)|b]$ given all the possible b values that B might take. The variance of $\mathrm{E}_{A|B}[f(A)|B]$ is the variance of this random variable. The variance of $\mathrm{E}_{A|b}[f(A)|b]$ in contrast is 0 since it is a fixed value. We will often be working with the random variables, $\mathrm{E}_{A|B}[f(A)|B]$ or $\mathrm{var}_{A|B}[f(A)|B]$, inside an expectation or variance: such as $\mathrm{var}_B[\mathrm{E}_{A|B}[f(A)|B]]$.

2.1.1 Law of total variance

The "law of total variance" can be written

$$\mathrm{var}_A[A] = \mathrm{var}_B[\mathrm{E}_{A|B}[A|B]] + \mathrm{E}_B[\mathrm{var}_{A|B}[A|B]]. \tag{4}$$

The subscripts on the inner expectations make it explicit that the expectations are being taken over the conditional distributions. $\mathrm{var}_{A|B}[A|B]$ and $\mathrm{E}_{A|B}[A|B]$ are random variables because the B in the conditional is a random variable. We take the expectation or variance with B fixed at one value, b, but B can take other values of b also.

Going forward, I will write the law or total variance more succinctly as

$$\mathrm{var}[A] = \mathrm{var}_B[\mathrm{E}[A|B]] + \mathrm{E}_B[\mathrm{var}[A|B]]. \tag{5}$$

I leave off the subscript on the inner conditional expectation or variance. Just remember that when you see a conditional in an expectation or variance, the integration is over over the conditional distribution of A conditioned on $B = b$. Even when you see $A|B$, the conditioning is on $B = b$ and the B indicates that this is a random variable because B can take different b values. When computing $\mathrm{var}_B[\mathrm{E}_{A|B}[A|B]]$, we will typically compute $\mathrm{E}_{A|b}[A|b]$ and then compute (or infer) the variance or expectation of that over all possible values of b.

The law of total variance will appear in this report in the following form:

$$\mathrm{var}_{XY}[f(\boldsymbol{Y}, \boldsymbol{X})] = \mathrm{var}_{Y^{(1)}}[\mathrm{E}_{XY|Y^{(1)}}[f(\boldsymbol{Y}, \boldsymbol{X})|\boldsymbol{Y}^{(1)}]] + \mathrm{E}_{Y^{(1)}}[\mathrm{var}_{XY|Y^{(1)}}[f(\boldsymbol{Y}, \boldsymbol{X})|\boldsymbol{Y}^{(1)}]], \tag{6}$$

where $f(\boldsymbol{Y}_t, \boldsymbol{X}_t)$ is some function of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ and $\boldsymbol{Y}^{(1)}$ is the observed data from $t = 1$ to T ($\boldsymbol{Y}^{(2)}$ is the unobserved data).

2.2 Model residuals conditioned on all the data

Define the smoothations $\widehat{\boldsymbol{v}}_t$ as:

$$\widehat{\boldsymbol{v}}_t = \boldsymbol{y}_t - \mathbf{Z}_t \widetilde{\boldsymbol{x}}_t^T - \mathbf{a}_t, \tag{7}$$

where $\widetilde{\boldsymbol{x}}_t^T$ is $\mathrm{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1)}]$. The smoothation is different from $\boldsymbol{v}_t$ because it uses $\widetilde{\boldsymbol{x}}_t^T$ not $\boldsymbol{x}_t$; $\boldsymbol{x}_t$ is not known, and $\widetilde{\boldsymbol{x}}_t^T$ is its estimate. $\widetilde{\boldsymbol{x}}_t^T$ is output by the Kalman smoother. $\boldsymbol{y}^{(1)}$ means all the observed data from $t = 1$ to T. $\boldsymbol{y}^{(1)}$ is a sample from the random variable $\boldsymbol{Y}^{(1)}$. The unobserved $\boldsymbol{y}$ will be termed $\boldsymbol{y}^{(2)}$ and is a sample from the random variable $\boldsymbol{Y}^{(2)}$. When $\boldsymbol{Y}$ appears without a superscript, it means both $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$ together. Similarly $\boldsymbol{y}$ means both $\boldsymbol{y}^{(1)}$ and $\boldsymbol{y}^{(2)}$ together—the observed data that we use to estimate $\widetilde{\boldsymbol{x}}_t^T$ and the unobserved data that we do not use and may or may not know. $\widehat{\boldsymbol{v}}_t$ exists for both $\boldsymbol{y}^{(1)}$ and $\boldsymbol{y}^{(2)}$, though we might not know $\boldsymbol{y}^{(2)}$ and thus might not know its corresponding $\widehat{\boldsymbol{v}}_t$. In some cases, however, we do know $\boldsymbol{y}^{(2)}$; they are data that we left out of our model fitting, in say a k-fold or leave-one-out cross-validation.

$\widehat{\boldsymbol{v}}_t$ is a sample from the random variable $\widehat{\boldsymbol{V}}_t$. We want to compute the mean and variance of this random variable over all possibles values that $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ might take. The mean of $\widehat{\boldsymbol{V}}_t$ is 0 and we are concerned only with computing the variance:

$$\mathrm{var}[\widehat{\boldsymbol{V}}_t] = \mathrm{var}_{XY}[\boldsymbol{Y}_t - \mathbf{Z}_t \mathrm{E}[\boldsymbol{X}_t|\boldsymbol{Y}^{(1)}] - \mathbf{a}_t]. \tag{8}$$

Notice we have an unconditional variance over $\boldsymbol{X},\boldsymbol{Y}$ on the outside and a conditional expectation over a specific value of $\boldsymbol{Y}^{(1)}$ on the inside (in the E[ ]).

From the law of total variance (Equation 4), we can write the variance of the model residuals as

$$\text{var}[\widehat{\boldsymbol{V}}_t] = \text{var}_{Y^{(1)}}[\,\text{E}[\widehat{\boldsymbol{V}}_t|\boldsymbol{Y}^{(1)}]] + \text{E}_{Y^{(1)}}[\,\text{var}[\widehat{\boldsymbol{V}}_t|\boldsymbol{Y}^{(1)}]]. \tag{9}$$

2.2.1 First term on right hand side of Equation 9

The random variable inside the var[ ] in the first term is

$$\text{E}[\widehat{\boldsymbol{V}}_t|\boldsymbol{Y}^{(1)}] = \text{E}[(\boldsymbol{Y}_t + \mathbf{Z}_t\,\text{E}[\boldsymbol{X}_t|\boldsymbol{Y}^{(1)}] + \mathbf{a}_t)|\boldsymbol{Y}^{(1)}]. \tag{10}$$

Let's consider this for a specific value $\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)}$.

$$\text{E}[\widehat{\boldsymbol{V}}_t|\boldsymbol{y}^{(1)}] = \text{E}[(\boldsymbol{Y}_t + \mathbf{Z}_t\,\text{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1)}] + \mathbf{a}_t)|\boldsymbol{y}^{(1)}] = \text{E}[\boldsymbol{Y}_t|\boldsymbol{y}^{(1)}] + \mathbf{Z}_t\,\text{E}[\,\text{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1)}]|\boldsymbol{y}^{(1)}] + \text{E}[\mathbf{a}_t|\boldsymbol{y}^{(1)}]. \tag{11}$$

$\text{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1)}]$ is a fixed value, and the expected value of a fixed value is itself. So $\text{E}[\,\text{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1)}]|\boldsymbol{y}^{(1)}] = \text{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1)}]$. Thus,

$$\text{E}[\widehat{\boldsymbol{V}}_t|\boldsymbol{y}^{(1)}] = \text{E}[\boldsymbol{Y}_t|\boldsymbol{y}^{(1)}] + \mathbf{Z}_t\,\text{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1)}] + \text{E}[\mathbf{a}_t|\boldsymbol{y}^{(1)}]. \tag{12}$$

We can move the conditional out and write

$$\text{E}[\widehat{\boldsymbol{V}}_t|\boldsymbol{y}^{(1)}] = \text{E}[(\boldsymbol{Y}_t + \mathbf{Z}_t\boldsymbol{X}_t + \mathbf{a}_t)|\boldsymbol{y}^{(1)}] = \text{E}[\boldsymbol{V}_t|\boldsymbol{y}^{(1)}]. \tag{13}$$

The right side is $\text{E}[\boldsymbol{V}_t|\boldsymbol{y}^{(1)}]$, no 'hat' on the $\boldsymbol{V}_t$, and this applies for all $\boldsymbol{y}^{(1)}$. This means that the first term in Equation 9 can be written with no hat on $\boldsymbol{V}$:

$$\text{var}_{Y^{(1)}}[\,\text{E}[\widehat{\boldsymbol{V}}_t|\boldsymbol{Y}^{(1)}]] = \text{var}_{Y^{(1)}}[\,\text{E}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}]]. \tag{14}$$

If $\boldsymbol{Y}_t$ were completely observed (no missing values), this would be zero since $\text{E}[\boldsymbol{V}_t|\boldsymbol{y}]$ would be a fixed value in that case. But $\boldsymbol{Y}_t$ is not assumed to be fully observed; it may have $\boldsymbol{Y}_t^{(2)}$ which is unobserved, or more precisely, not included in the estimation of VV_t for whatever reason ('unknown" because it was unobserved being one reason). The derivation of $\text{E}[\boldsymbol{Y}_t|\boldsymbol{y}^{(1)}]$ is given in Holmes (2012)[4].

Using the law of total variance, we can re-write $\text{var}[\boldsymbol{V}_t]$ as:

$$\text{var}[\boldsymbol{V}_t] = \text{var}_{Y^{(1)}}[\,\text{E}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}]] + \text{E}_{Y^{(1)}}[\,\text{var}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}]]. \tag{15}$$

From Equation 15, we can solve for $\text{var}_{Y^{(1)}}[\,\text{E}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}]]$:

$$\text{var}_{Y^{(1)}}[\,\text{E}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}]] = \text{var}[\boldsymbol{V}_t] - \text{E}_{Y^{(1)}}[\,\text{var}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}]]. \tag{16}$$

From Equation 3, we know that $\text{var}[\boldsymbol{V}_t] = \mathbf{R}_t$ (this is the unconditional variance). Thus,

$$\text{var}_{Y^{(1)}}[\,\text{E}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}]] = \mathbf{R}_t - \text{E}_{Y^{(1)}}[\,\text{var}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}]]. \tag{17}$$

The second term in Equation 17 to the right of the equal sign and inside the expectation is $\text{var}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}]$. This is the variance of $\boldsymbol{V}_t$ with $\boldsymbol{Y}^{(1)}$ held at a specific fixed $\boldsymbol{y}^{(1)}$. The variability in $\text{var}[\boldsymbol{V}_t|\boldsymbol{y}^{(1)}]$ (notice $\boldsymbol{y}^{(1)}$ not $\boldsymbol{Y}^{(1)}$ now) comes from $\boldsymbol{X}_t$ and $\boldsymbol{Y}^{(2)}$ which are random variables. Let's compute this variance for a specific $\boldsymbol{y}^{(1)}$ value.

$$\text{var}[\boldsymbol{V}_t|\boldsymbol{y}^{(1)}] = \text{var}[\boldsymbol{Y}_t - \mathbf{Z}_t\boldsymbol{X}_t - \mathbf{a}_t|\boldsymbol{y}^{(1)}]. \tag{18}$$

Notice that there is no E (expectation) on the $\boldsymbol{X}_t$; this is $\boldsymbol{V}_t$ not $\widehat{\boldsymbol{V}}_t$. $\mathbf{a}_t$ is a fixed value and can be dropped.

Equation 18 can be written as[5]:

$$\begin{aligned}
\text{var}[\boldsymbol{V}_t|\boldsymbol{y}^{(1)}] &= \text{var}[\boldsymbol{Y}_t - \mathbf{Z}_t\boldsymbol{X}_t|\boldsymbol{y}^{(1)}] \\
&= \text{var}[-\mathbf{Z}_t\boldsymbol{X}_t|\boldsymbol{y}^{(1)}] + \text{var}[\boldsymbol{Y}_t|\boldsymbol{y}^{(1)}] + \text{cov}[\boldsymbol{Y}_t, -\mathbf{Z}_t\boldsymbol{X}_t|\boldsymbol{y}^{(1)}] + \text{cov}[-\mathbf{Z}_t\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{y}^{(1)}] \\
&= \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T\mathbf{Z}_t^\top + \widetilde{\mathbf{U}}_t^T - \widetilde{\mathbf{S}}_t^T\mathbf{Z}_t^\top - \mathbf{Z}_t(\widetilde{\mathbf{S}}_t^T)^\top.
\end{aligned} \tag{19}$$

[4] $\text{E}[\boldsymbol{Y}_t^{(2)}|\boldsymbol{y}^{(1)}]$ is not $\mathbf{Z}_t\widetilde{\boldsymbol{x}}_t^T + \mathbf{a}_t$ in general since $\boldsymbol{Y}_t^{(2)}$ and $\boldsymbol{Y}_t^{(1)}$ may be correlated through $\mathbf{R}$ in addition to being correlated through $\widetilde{\boldsymbol{x}}_t^T$

[5] $\text{var}(A+B) = \text{var}(A) + \text{var}(B) + \text{cov}(A,B) + \text{cov}(B,A)$

$\widetilde{\boldsymbol{V}}_t^T = \text{var}[\boldsymbol{X}_t|\boldsymbol{y}^{(1)}]$ and is output by the Kalman smoother. $\widetilde{\mathbf{U}}_t^T = \text{var}[\boldsymbol{Y}_t|\boldsymbol{y}^{(1)}]$ and $\widetilde{\mathbf{S}}_t^T = \text{cov}[\boldsymbol{Y}_t, \boldsymbol{X}_t|\boldsymbol{y}^{(1)}]$. The equations for these are given in Holmes (2012) and are output by the `MARSShatyt()` function in the {MARSS} package[6]. If there were no missing data, i.e., if $\boldsymbol{y}^{(1)} = \boldsymbol{y}$, then $\widetilde{\mathbf{U}}_t^T$ and $\widetilde{\mathbf{S}}_t^T$ would be zero because $\boldsymbol{Y}_t$ would be fixed at $\boldsymbol{y}_t$. This would reduce Equation 19 to $\mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T\mathbf{Z}_t^\top$. But we are concerned with the case where there are missing values. Those missing values need not be for all t. That is, there may be some observed y at time t and some missing y. $\boldsymbol{y}_t$ is multivariate.

From Equation 19, we know $\text{var}[\boldsymbol{V}_t|\boldsymbol{y}^{(1)}]$ for a specific $\boldsymbol{y}^{(1)}$. We want $\text{E}_{Y^{(1)}}[\text{var}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}]]$ which is its expected value over all possible values of $\boldsymbol{y}^{(1)}$. $\boldsymbol{V}_t$ is a multivariate normal random variable with two random variables $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$. The conditional variance of a multivariate Normal does not depend on the value that you are conditioning on. Let the $\mathbf{A}$ be a N-dimensional multivariate Normal random variable partitioned into $\mathbf{A}_1$ and $\mathbf{A}_2$ with variance-covariance matrix $\Sigma = \begin{bmatrix}\Sigma_1 & \Sigma_{12}\\ \Sigma_2 1 & \Sigma_2\end{bmatrix}$. The variance-covariance matrix of $\mathbf{A}$ conditioned on $\mathbf{A}_1 = \mathbf{a}$ is $\Sigma = \begin{bmatrix}0 & 0\\ 0 & \Sigma_2 - \Sigma_{12}\Sigma_1\Sigma_{21}\end{bmatrix}$. Notice that $\mathbf{a}$ does not appear in the conditional variance matrix. This means that $\text{var}[\boldsymbol{V}_t|\boldsymbol{y}^{(1)}]$ does not depend on $\boldsymbol{y}^{(1)}$. Its variance only depends on the MARSS model parameters[7].

Because $\text{var}[\boldsymbol{V}_t|\boldsymbol{y}^{(1)}]$ only depends on the MARSS parameters values, $\mathbf{Q}$, $\mathbf{B}$, $\mathbf{R}$, etc., the second term in Equation 16, $\text{E}_{Y^{(1)}}[\text{var}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}]]$, is equal to $\text{var}[\boldsymbol{V}_t|\boldsymbol{y}^{(1)}]$ (Equation 19). Putting this into Equation 17, we have

$$\text{var}_{Y^{(1)}}[\text{E}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)})]] = \mathbf{R}_t - \text{var}[\boldsymbol{V}_t|\boldsymbol{y}^{(1)}] = \mathbf{R}_t - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T\mathbf{Z}_t^\top - \widetilde{\mathbf{U}}_t^T + \widetilde{\mathbf{S}}_t^T\mathbf{Z}_t^\top + \mathbf{Z}_t(\widetilde{\mathbf{S}}_t^T)^\top. \tag{20}$$

Since $\text{var}_{Y^{(1)}}[\text{E}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)})]] = \text{var}_{Y^{(1)}}[\text{E}[\widehat{\boldsymbol{V}}_t|\boldsymbol{Y}^{(1)})]]$ (Equation 14), this means that the first term in Equation 9 is

$$\text{var}_{Y^{(1)}}[\text{E}[\widehat{\boldsymbol{V}}_t|\boldsymbol{Y}^{(1)})]] = \mathbf{R}_t - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T\mathbf{Z}_t^\top - \widetilde{\mathbf{U}}_t^T + \widetilde{\mathbf{S}}_t^T\mathbf{Z}_t^\top + \mathbf{Z}_t(\widetilde{\mathbf{S}}_t^T)^\top. \tag{21}$$

2.2.2 Second term on right hand side of Equation 9

Consider the second term in Equation 9. This term is

$$\text{E}_{Y^{(1)}}[\text{var}[\widehat{\boldsymbol{V}}_t|\boldsymbol{Y}^{(1)}]] = \text{E}_{Y^{(1)}}[\text{var}[(\boldsymbol{Y}_t - \mathbf{Z}_t\,\text{E}[\boldsymbol{X}_t|\boldsymbol{Y}^{(1)}] - \mathbf{a}_t)|\boldsymbol{Y}^{(1)}]]. \tag{22}$$

The middle term is:

$$\text{E}_{Y^{(1)}}[\text{var}[\text{E}[\boldsymbol{X}_t|\boldsymbol{Y}^{(1)}]|\boldsymbol{Y}^{(1)}]]. \tag{23}$$

Let's solve the inner part for a specific $\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)}$. $\text{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1)}]$ is a fixed value. Thus $\text{var}[\text{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1)}]|\boldsymbol{y}^{(1)}] = 0$ since the variance of a fixed value is 0. This is true for all $\boldsymbol{y}^{(1)}$ so the middle term reduces to 0. $\mathbf{a}_t$ is also fixed and its variance is also 0. The covariance between a random variable and a fixed value is 0[8]. Thus for a specific $\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)}$, the inside of the right hand side expectation reduces to $\text{var}[\boldsymbol{Y}_t|\boldsymbol{y}^{(1)}]$ which is $\widetilde{\mathbf{U}}_t^T$. As noted in the previous section, $\widetilde{\mathbf{U}}_t^T$ is only a function of the MARSS parameters; it is not a function of $\boldsymbol{y}^{(1)}$ and $\text{var}[\boldsymbol{Y}_t|\boldsymbol{y}^{(1)}] = \widetilde{\mathbf{U}}_t^T$ for all $\boldsymbol{y}^{(1)}$. Thus the second term in Equation 9 is simply $\widetilde{\mathbf{U}}_t^T$:

$$\text{E}_{Y^{(1)}}[\text{var}[\widehat{\boldsymbol{V}}_t|\boldsymbol{Y}^{(1)}]] = \text{var}[\widehat{\boldsymbol{V}}_t|\boldsymbol{y}^{(1)}] = \widetilde{\mathbf{U}}_t^T. \tag{24}$$

2.2.3 Putting together the first and second terms

We can now put the first and second terms in Equation 9 together (Equations 21 and 24) and write out the variance of the model residuals:

$$\begin{aligned}\text{var}[\widehat{\boldsymbol{V}}_t] &= \mathbf{R}_t - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T\mathbf{Z}_t^\top - \widetilde{\mathbf{U}}_t^T + \widetilde{\mathbf{S}}_t^T\mathbf{Z}_t^\top + \mathbf{Z}_t(\widetilde{\mathbf{S}}_t^T)^\top + \widetilde{\mathbf{U}}_t^T\\ &= \mathbf{R}_t - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T\mathbf{Z}_t^\top + \widetilde{\mathbf{S}}_t^T\mathbf{Z}_t^\top + \mathbf{Z}_t(\widetilde{\mathbf{S}}_t^T)^\top.\end{aligned} \tag{25}$$

[6] $\widetilde{\mathbf{U}}_t^T$ is `OtT - tcrossprod(ytT)` in the `MARSShatyt()` output.

[7] This also implies that $\widetilde{\boldsymbol{V}}_t^T$, $\widetilde{\mathbf{U}}_t^T$ and $\widetilde{\mathbf{S}}_t^T$ do not depend on $\boldsymbol{y}^{(1)}$, which is rather counter-intuitive since $\boldsymbol{y}^{(1)}$ is part of the algorithm (Kalman filter and smoother) used to compute these values. The value of $\boldsymbol{y}^{(1)}$ affects the expected values of $\boldsymbol{X}_t$ but only its presence (versus absence) affects $\boldsymbol{X}_t$'s conditional variance.

[8] $\text{var}[A + B] = \text{var}[A] + \text{var}[B] + \text{cov}[A, B]$

Equation 25 will reduce to $\mathbf{R}_t - \mathbf{Z}_t \widetilde{\boldsymbol{V}}_t^T \mathbf{Z}_t^\top$ if $\boldsymbol{y}_t$ has no missing values since $\widetilde{\mathbf{S}}_t^T = 0$ in this case. If $\boldsymbol{y}_t$ is all missing values, $\widetilde{\mathbf{S}}_t^T = \mathbf{Z}_t \widetilde{\boldsymbol{V}}_t^T$ because

$$\text{cov}[\boldsymbol{Y}_t, \boldsymbol{X}_t | \boldsymbol{y}^{(1)}] = \text{cov}[\mathbf{Z}_t \boldsymbol{X}_t + \mathbf{a}_t + \boldsymbol{V}_t, \boldsymbol{X}_t | \boldsymbol{y}^{(1)}] = \text{cov}[\mathbf{Z}_t \boldsymbol{X}_t, \boldsymbol{X}_t | \boldsymbol{y}^{(1)}] = \mathbf{Z}_t \,\text{cov}[\boldsymbol{X}_t, \boldsymbol{X}_t | \boldsymbol{y}^{(1)}] = \mathbf{Z}_t \widetilde{\boldsymbol{V}}_t^T. \quad (26)$$

The reduction in Equation 26 occurs because $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ and by extension $\boldsymbol{V}_t$ and $\boldsymbol{X}_t$ are independent in the form of MARSS model used in this report (Equation 1)[9]. Thus when $\boldsymbol{y}_t$ is all missing values, Equation 25 will reduce to $\mathbf{R}_t + \mathbf{Z}_t \widetilde{\boldsymbol{V}}_t^T \mathbf{Z}_t^\top$. The behavior if $\boldsymbol{y}_t$ has some missing and some not missing values depends on whether $\mathbf{R}_t$ is a diagonal matrix or not, i.e., if the $\boldsymbol{y}_t^{(1)}$ and $\boldsymbol{y}_t^{(2)}$ are correlated.

2.3 State residuals conditioned on the data

The state residuals are $\boldsymbol{x}_t - (\mathbf{B}_t \boldsymbol{x}_{t-1} + \mathbf{u}_t) = \boldsymbol{w}_t$. The unconditional expected value of the state residuals is $\text{E}[\boldsymbol{X}_t - (\mathbf{B}_t \boldsymbol{X}_{t-1} + \mathbf{u}_t)] = \text{E}[\boldsymbol{W}_t] = 0$ and the unconditional variance of the state residuals is

$$\text{var}[\boldsymbol{X}_t - (\mathbf{B}_t \boldsymbol{X}_{t-1} + \mathbf{u}_t)] = \text{var}[\boldsymbol{W}_t] = \mathbf{Q}_t \quad (27)$$

based on the definition of $\boldsymbol{W}_t$ in Equation 1. The conditional state residuals (conditioned on the data from $t = 1$ to $t = T$) are defined as

$$\widehat{\boldsymbol{w}}_t = \widetilde{\boldsymbol{x}}_t^T - \mathbf{B}_t \widetilde{\boldsymbol{x}}_{t-1}^T - \mathbf{u}_t. \quad (28)$$

where $\widetilde{\boldsymbol{x}}_t^T = E[\boldsymbol{X}_t | \boldsymbol{y}^{(1)}]$ and $\widetilde{\boldsymbol{x}}_{t-1}^T = E[\boldsymbol{X}_{t-1} | \boldsymbol{y}^{(1)}]$. $\widehat{\boldsymbol{w}}_t$ is a sample from the random variable $\widehat{\boldsymbol{W}}_t$; random over different possible data sets. The expected value of $\widehat{\boldsymbol{W}}_t$ is 0, and we are concerned with computing its variance.

We can write the variance of $\boldsymbol{W}_t$ (no hat) using the law of total variance.

$$\text{var}[\boldsymbol{W}_t] = \text{var}_{Y^{(1)}}[\,\text{E}[\boldsymbol{W}_t | \boldsymbol{Y}^{(1)}]] + \text{E}_{Y^{(1)}}[\,\text{var}[\boldsymbol{W}_t | \boldsymbol{Y}^{(1)}]]. \quad (29)$$

Notice that

$$\text{E}[\boldsymbol{W}_t | \boldsymbol{y}^{(1)}] = \text{E}[(\boldsymbol{X}_t - \mathbf{B}_t \boldsymbol{X}_{t-1} - \mathbf{u}_t) | \boldsymbol{y}^{(1)}] = \widetilde{\boldsymbol{x}}_t^T - \mathbf{B}_t \widetilde{\boldsymbol{x}}_{t-1}^T - \mathbf{u}_t = \text{E}[\widehat{\boldsymbol{W}}_t | \boldsymbol{y}^{(1)}] = \widehat{\boldsymbol{w}}_t. \quad (30)$$

This is true for all $\boldsymbol{y}^{(1)}$, thus $\text{E}[\boldsymbol{W}_t | \boldsymbol{Y}^{(1)}]$ is $\widehat{\boldsymbol{W}}_t$, and $\text{var}_{Y^{(1)}}[\,\text{E}[\boldsymbol{W}_t | \boldsymbol{Y}^{(1)}]] = \text{var}[\widehat{\boldsymbol{W}}_t]$. Equation 29 can thus be written

$$\text{var}[\boldsymbol{W}_t] = \text{var}[\widehat{\boldsymbol{W}}_t] + \text{E}_{Y^{(1)}}[\,\text{var}[\boldsymbol{W}_t | \boldsymbol{Y}^{(1)}]]. \quad (31)$$

Solve for $\text{var}[\widehat{\boldsymbol{W}}_t]$:

$$\text{var}[\widehat{\boldsymbol{W}}_t] = \text{var}[\boldsymbol{W}_t] - \text{E}_{Y^{(1)}}[\,\text{var}[\boldsymbol{W}_t | \boldsymbol{Y}^{(1)}]]. \quad (32)$$

The variance in the expectation on the far right for a specific $\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)}$ is

$$\text{var}[\boldsymbol{W}_t | \boldsymbol{y}^{(1)}] = \text{var}[(\boldsymbol{X}_t - \mathbf{B}_t \boldsymbol{X}_{t-1} - \mathbf{u}_t) | \boldsymbol{y}^{(1)}]. \quad (33)$$

$\mathbf{u}_t$ is not a random variable and can be dropped. Thus[10],

$$\begin{aligned} \text{var}[\boldsymbol{W}_t | \boldsymbol{y}^{(1)}] &= \text{var}[(\boldsymbol{X}_t - \mathbf{B}_t \boldsymbol{X}_{t-1}) | \boldsymbol{y}^{(1)}] \\ &= \text{var}[\boldsymbol{X}_t | \boldsymbol{y}^{(1)}] + \text{var}[\mathbf{B}_t \boldsymbol{X}_{t-1} | \boldsymbol{y}^{(1)}] + \text{cov}[\boldsymbol{X}_t, -\mathbf{B}_t \boldsymbol{X}_{t-1} | \boldsymbol{y}^{(1)}] + \text{cov}[-\mathbf{B}_t \boldsymbol{X}_{t-1}, \boldsymbol{X}_t | \boldsymbol{y}^{(1)}] \\ &= \widetilde{\boldsymbol{V}}_t^T + \mathbf{B}_t \widetilde{\boldsymbol{V}}_{t-1}^T \mathbf{B}_t^\top - \widetilde{\boldsymbol{V}}_{t,t-1}^T \mathbf{B}_t^\top - \mathbf{B}_t \widetilde{\boldsymbol{V}}_{t-1,t}^T. \end{aligned} \quad (34)$$

Again this is conditional multivariate normal variance, and its value does not depend on the value, $\boldsymbol{y}^{(1)}$ that we are conditioning on. It depends only on the parameters values, $\mathbf{Q}$, $\mathbf{B}$, $\mathbf{R}$, etc., and is the same for all values of $\boldsymbol{y}^{(1)}$. So $\text{E}_{Y^{(1)}}[\,\text{var}[\boldsymbol{W}_t | \boldsymbol{Y}^{(1)}]] = \text{var}[\boldsymbol{W}_t | \boldsymbol{y}^{(1)}]$, using any value of $\boldsymbol{y}^{(1)}$. Thus

$$\text{E}_{Y^{(1)}}[\,\text{var}[\boldsymbol{W}_t | \boldsymbol{Y}^{(1)}]] = \widetilde{\boldsymbol{V}}_t^T + \mathbf{B}_t \widetilde{\boldsymbol{V}}_{t-1}^T \mathbf{B}_t^\top - \widetilde{\boldsymbol{V}}_{t,t-1}^T \mathbf{B}_t^\top - \mathbf{B}_t \widetilde{\boldsymbol{V}}_{t-1,t}^T. \quad (35)$$

Putting $\text{E}_{Y^{(1)}}[\,\text{var}[\boldsymbol{W}_t | \boldsymbol{Y}^{(1)}]]$ from Equation 35 and $\text{var}[\boldsymbol{W}_t] = \mathbf{Q}_t$ into Equation 32, the variance of the conditional state residuals is

$$\text{var}[\widehat{\boldsymbol{W}}_t] = \mathbf{Q}_t - \widetilde{\boldsymbol{V}}_t^T - \mathbf{B}_t \widetilde{\boldsymbol{V}}_{t-1}^T \mathbf{B}_t^\top + \widetilde{\boldsymbol{V}}_{t,t-1}^T \mathbf{B}_t^\top + \mathbf{B}_t \widetilde{\boldsymbol{V}}_{t-1,t}^T. \quad (36)$$

[9] This is not the case for the Harvey et al. (1998) form of the MARSS model where $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ are allowed to be correlated.

[10] $\text{var}[A - B] = \text{var}[A] + \text{var}[B] + \text{cov}[A, -B] + \text{cov}[-B, A]$. Be careful about the signs in this case as they are a little non-intuitive.

2.4 Covariance of the conditional model and state residuals

The unconditional model and state residuals, $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$, are independent by definition[11] (in Equation 1), i.e., $\mathrm{cov}[\boldsymbol{V}_t, \boldsymbol{W}_t] = 0$. However the conditional model and state residuals, $\mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_t]$, are not independent since both depend on $\boldsymbol{y}^{(1)}$. Using the law of total covariance, we can write

$$\mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_t] = \mathrm{cov}_{Y^{(1)}}[\mathrm{E}[\widehat{\boldsymbol{V}}_t|\boldsymbol{Y}^{(1)}], \mathrm{E}[\widehat{\boldsymbol{W}}_t|\boldsymbol{Y}^{(1)}]] + \mathrm{E}_{Y^{(1)}}[\mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_t|\boldsymbol{Y}^{(1)}]]. \tag{37}$$

For a specific value of $\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)}$, the covariance in the second term on the right is $\mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_t|\boldsymbol{y}^{(1)}]$. Conditioned on a specific value of $\boldsymbol{Y}^{(1)}$, $\widehat{\boldsymbol{W}}_t$ is a fixed value, $\widehat{\boldsymbol{w}}_t = \widetilde{\boldsymbol{x}}_t^T - \mathbf{B}_t\widetilde{\boldsymbol{x}}_{t-1}^T - \mathbf{u}_t$, and conditioned on $\boldsymbol{y}^{(1)}$, $\widetilde{\boldsymbol{x}}_t^T$ and $\widetilde{\boldsymbol{x}}_{t-1}^T$ are fixed values. $\mathbf{u}_t$ is also fixed; it is a parameter. $\widehat{\boldsymbol{V}}_t$ is not a fixed value because it has $\boldsymbol{Y}_t^{(2)}$ and that is a random variable. Thus $\mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_t|\boldsymbol{y}^{(1)}]$ is the covariance between a random variable and a fixed variable and thus the covariance is 0. This is true for all $\boldsymbol{y}^{(1)}$. Thus the second right-side term in Equation 37 is zero, and the equation reduces to

$$\mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_t] = \mathrm{cov}_{Y^{(1)}}[\mathrm{E}[\widehat{\boldsymbol{V}}_t|\boldsymbol{Y}^{(1)}], \mathrm{E}[\widehat{\boldsymbol{W}}_t|\boldsymbol{Y}^{(1)}]]. \tag{38}$$

Notice that $\mathrm{E}[\widehat{\boldsymbol{W}}_t|\boldsymbol{y}^{(1)}] = \mathrm{E}[\boldsymbol{W}_t|\boldsymbol{y}^{(1)}]$ and $\mathrm{E}[\widehat{\boldsymbol{V}}_t|\boldsymbol{y}^{(1)}] = \mathrm{E}[\boldsymbol{V}_t|\boldsymbol{y}^{(1)}]$ since

$$\mathrm{E}[\boldsymbol{W}_t|\boldsymbol{y}^{(1)}] = \mathrm{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1)}] - \mathbf{B}_t\,\mathrm{E}[\boldsymbol{X}_{t-1}|\boldsymbol{y}^{(1)}] - \mathbf{u}_t = \widetilde{\boldsymbol{x}}_t^T - \mathbf{B}_t\widetilde{\boldsymbol{x}}_{t-1}^T - \mathbf{u}_t = \widehat{\boldsymbol{w}}_t = \mathrm{E}[\widehat{\boldsymbol{W}}_t|\boldsymbol{y}^{(1)}] \tag{39}$$

and

$$\mathrm{E}[\boldsymbol{V}_t|\boldsymbol{y}^{(1)}] = \mathrm{E}[\boldsymbol{Y}_t|\boldsymbol{y}^{(1)}] - \mathbf{Z}_t\,\mathrm{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1)}] - \mathbf{a}_t = \mathrm{E}[\boldsymbol{Y}_t|\boldsymbol{y}^{(1)}] - \mathbf{Z}_t\widetilde{\boldsymbol{x}}_t^T - \mathbf{a}_t = \mathrm{E}[\widehat{\boldsymbol{V}}_t|\boldsymbol{y}^{(1)}]. \tag{40}$$

Thus the right side of Equation 38 can be written in terms of $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ instead of $\widehat{\boldsymbol{V}}_t$ and $\widehat{\boldsymbol{W}}_t$:

$$\mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_t] = \mathrm{cov}_{Y^{(1)}}[\mathrm{E}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}], \mathrm{E}[\boldsymbol{W}_t|\boldsymbol{Y}^{(1)}]]. \tag{41}$$

Using the law of total covariance, we can write:

$$\mathrm{cov}[\boldsymbol{V}_t, \boldsymbol{W}_t] = \mathrm{E}_{Y^{(1)}}[\mathrm{cov}[\boldsymbol{V}_t, \boldsymbol{W}_t|\boldsymbol{Y}^{(1)}]] + \mathrm{cov}_{Y^{(1)}}[\mathrm{E}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}], \mathrm{E}[\boldsymbol{W}_t|\boldsymbol{Y}^{(1)}]]. \tag{42}$$

The unconditional covariance of $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ is 0. Thus the left side of Equation 42 is 0 and we can rearrange the equation as

$$\mathrm{cov}_{Y^{(1)}}[\mathrm{E}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}], \mathrm{E}[\boldsymbol{W}_t|\boldsymbol{Y}^{(1)}]] = -\mathrm{E}_{Y^{(1)}}[\mathrm{cov}[\boldsymbol{V}_t, \boldsymbol{W}_t|\boldsymbol{Y}^{(1)}]]. \tag{43}$$

Combining Equation 41 and 43, we get

$$\mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_t] = -\mathrm{E}_{Y^{(1)}}[\mathrm{cov}[\boldsymbol{V}_t, \boldsymbol{W}_t|\boldsymbol{Y}^{(1)}]], \tag{44}$$

and our problem reduces to solving for the conditional covariance of the model and state residuals (right side of Equation 44).

For a specific $\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)}$, the conditional covariance $\mathrm{cov}[\boldsymbol{V}_t, \boldsymbol{W}_t|\boldsymbol{y}^{(1)}]$ can be written out as

$$\mathrm{cov}[\boldsymbol{V}_t, \boldsymbol{W}_t|\boldsymbol{y}^{(1)}] = \mathrm{cov}[\boldsymbol{Y}_t - \mathbf{Z}_t\boldsymbol{X}_t - \mathbf{a}_t,\, \boldsymbol{X}_t - \mathbf{B}_t\boldsymbol{X}_{t-1} - \mathbf{u}_t|\boldsymbol{y}^{(1)}]. \tag{45}$$

$\mathbf{a}_t$ and $\mathbf{u}_t$ are fixed values and can be dropped. Thus[12]

$$\begin{aligned}\mathrm{cov}[\boldsymbol{V}_t, \boldsymbol{W}_t|\boldsymbol{y}^{(1)}] &= \mathrm{cov}[\boldsymbol{Y}_t - \mathbf{Z}_t\boldsymbol{X}_t, \boldsymbol{X}_t - \mathbf{B}_t\boldsymbol{X}_{t-1}|\boldsymbol{y}^{(1)}] \\ &= \mathrm{cov}[\boldsymbol{Y}_t, \boldsymbol{X}_t|\boldsymbol{y}^{(1)}] + \mathrm{cov}[\boldsymbol{Y}_t, -\mathbf{B}_t\boldsymbol{X}_{t-1}|\boldsymbol{y}^{(1)}] + \mathrm{cov}[-\mathbf{Z}_t\boldsymbol{X}_t, \boldsymbol{X}_t|\boldsymbol{y}^{(1)}] + \mathrm{cov}[-\mathbf{Z}_t\boldsymbol{X}_t, -\mathbf{B}_t\boldsymbol{X}_{t-1}|\boldsymbol{y}^{(1)}] \\ &= \widetilde{\mathbf{S}}_t^T - \widetilde{\mathbf{S}}_{t,t-1}^T\mathbf{B}_t^\top - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T + \mathbf{Z}_t\widetilde{\boldsymbol{V}}_{t,t-1}^T\mathbf{B}_t^\top,\end{aligned} \tag{46}$$

where $\widetilde{\mathbf{S}}_t^T = \mathrm{cov}[\boldsymbol{Y}_t, \boldsymbol{X}_t|\boldsymbol{y}^{(1)}]$ and $\widetilde{\mathbf{S}}_{t,t-1}^T = \mathrm{cov}[\boldsymbol{Y}_t, \boldsymbol{X}_{t-1}|\boldsymbol{y}^{(1)}]$; the equations for $\widetilde{\mathbf{S}}_t^T$ and $\widetilde{\mathbf{S}}_{t,t-1}^T$ are given in Holmes (2012) and are output by the `MARSShatyt()` function in the {MARSS} package.

[11] This independence is specific to the way the MARSS model for this report (Equation 1). It is possible for the model and state residuals to covary. In the MARSS model written in Harvey et al. (1998) form, they do covary.

[12] $\mathrm{cov}[\mathbf{BA}, \mathbf{CD}] = \mathbf{B}\,\mathrm{cov}[\mathbf{A}, \mathbf{D}]\mathbf{C}^\top$.

Both $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ are multivariate normal random variables that depend on $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$ and the conditional covariance is not a function of the variable that we condition on (in this case $\boldsymbol{y}^{(1)}$). The conditional covariance is only a function of the MARSS parameters[13]. Thus

$$\mathrm{E}_{Y^{(1)}}[\,\mathrm{cov}[\boldsymbol{V}_t, \boldsymbol{W}_t|\boldsymbol{Y}^{(1)}]] = \mathrm{cov}[\boldsymbol{V}_t, \boldsymbol{W}_t|\boldsymbol{y}^{(1)}] = \widetilde{\mathbf{S}}_t^T - \widetilde{\mathbf{S}}_{t,t-1}^T\mathbf{B}_t^\top - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T + \mathbf{Z}_t\widetilde{\boldsymbol{V}}_{t,t-1}^T\mathbf{B}_t^\top. \tag{47}$$

$\mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_t]$ is the negative of this (Equation 44), thus

$$\mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_t] = -\widetilde{\mathbf{S}}_t^T + \widetilde{\mathbf{S}}_{t,t-1}^T\mathbf{B}_t^\top + \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_{t,t-1}^T\mathbf{B}_t^\top. \tag{48}$$

The Harvey et al. algorithm (next section) gives the joint distribution of the model residuals at time t and state residuals at time $t+1$. Using the law of total covariance as above, the covariance in this case is

$$\mathrm{cov}_{Y^{(1)}}[\,\mathrm{E}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}],\, \mathrm{E}[\boldsymbol{W}_{t+1}|\boldsymbol{Y}^{(1)}]] = -\,\mathrm{E}_{Y^{(1)}}[\,\mathrm{cov}[\boldsymbol{V}_t, \boldsymbol{W}_{t+1}|\boldsymbol{Y}^{(1)}]] \tag{49}$$

and

$$\begin{aligned}\mathrm{cov}[\boldsymbol{V}_t, \boldsymbol{W}_{t+1}|\boldsymbol{y}^{(1)}] &= \mathrm{cov}[\boldsymbol{Y}_t - \mathbf{Z}_t\boldsymbol{X}_t - \mathbf{a}_t,\, \boldsymbol{X}_{t+1} - \mathbf{B}_{t+1}\boldsymbol{X}_t - \mathbf{u}_{t+1}|\boldsymbol{y}^{(1)}] \\ &= \mathrm{cov}[\boldsymbol{Y}_t - \mathbf{Z}_t\boldsymbol{X}_t,\, \boldsymbol{X}_{t+1} - \mathbf{B}_{t+1}\boldsymbol{X}_t|\boldsymbol{y}^{(1)}] \\ &= \widetilde{\mathbf{S}}_{t,t+1}^T - \widetilde{\mathbf{S}}_t^T\mathbf{B}_{t+1}^\top - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_{t,t+1}^T + \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T\mathbf{B}_{t+1}^\top.\end{aligned} \tag{50}$$

Thus,

$$\begin{aligned}\mathrm{cov}_{Y^{(1)}}[\,\mathrm{E}[\boldsymbol{V}_t|\boldsymbol{Y}^{(1)}],\, \mathrm{E}[\boldsymbol{W}_{t+1}|\boldsymbol{Y}^{(1)}]] &= -\,\mathrm{E}_{Y^{(1)}}[\,\mathrm{cov}[\boldsymbol{V}_t, \boldsymbol{W}_{t+1}|\boldsymbol{Y}^{(1)}]] \\ &= -\widetilde{\mathbf{S}}_{t,t+1}^T + \widetilde{\mathbf{S}}_t^T\mathbf{B}_{t+1}^\top + \mathbf{Z}_t\widetilde{\boldsymbol{V}}_{t,t+1}^T - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T\mathbf{B}_{t+1}^\top.\end{aligned} \tag{51}$$

2.5 Joint distribution of the conditional residuals

We now can write the variance of the joint distribution of the conditional residuals. Define

$$\widehat{\varepsilon}_t = \begin{bmatrix}\widehat{\boldsymbol{v}}_t \\ \widehat{\boldsymbol{w}}_t\end{bmatrix} = \begin{bmatrix}\boldsymbol{y}_t - \mathbf{Z}_t\widetilde{\boldsymbol{x}}_t^T - \mathbf{a}_t \\ \widetilde{\boldsymbol{x}}_t^T - \mathbf{B}_t\widetilde{\boldsymbol{x}}_{t-1}^T - \mathbf{u}_t\end{bmatrix}. \tag{52}$$

$\widehat{\varepsilon}_t$ is a sample drawn from the distribution of the random variable $\widehat{\mathcal{E}}_t$. The expected value of $\widehat{\mathcal{E}}_t$ over all possible $\boldsymbol{y}$ is 0 and the variance of $\widehat{\mathcal{E}}_t$ is

$$\widehat{\Sigma}_t = \mathrm{var}[\widehat{\mathcal{E}}_t] = \left[\begin{array}{c|c}\mathrm{var}[\widehat{\boldsymbol{V}}_t] & \mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_t] \\ \hline (\,\mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_t])^\top & \mathrm{var}[\widehat{\boldsymbol{W}}_t]\end{array}\right] \tag{53}$$

which is

$$\left[\begin{array}{c|c}\mathbf{R}_t - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T\mathbf{Z}_t^\top + \widetilde{\mathbf{S}}_t^T\mathbf{Z}_t^\top + \mathbf{Z}_t(\widetilde{\mathbf{S}}_t^T)^\top & -\widetilde{\mathbf{S}}_t^T + \widetilde{\mathbf{S}}_{t,t-1}^T\mathbf{B}_t^\top + \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_{t,t-1}^T\mathbf{B}_t^\top \\ \hline (-\widetilde{\mathbf{S}}_t^T + \widetilde{\mathbf{S}}_{t,t-1}^T\mathbf{B}_t^\top + \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^T - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_{t,t-1}^T\mathbf{B}_t^\top)^\top & \mathbf{Q}_t - \widetilde{\boldsymbol{V}}_t^T - \mathbf{B}_t\widetilde{\boldsymbol{V}}_{t-1}^T\mathbf{B}_t^\top + \widetilde{\boldsymbol{V}}_{t,t-1}^T\mathbf{B}_t^\top + \mathbf{B}_t\widetilde{\boldsymbol{V}}_{t-1,t}^T\end{array}\right], \tag{54}$$

where $\widetilde{\mathbf{S}}_t^T = \mathrm{cov}[\boldsymbol{Y}_t, \boldsymbol{X}_t|\boldsymbol{y}^{(1)}]$, $\widetilde{\mathbf{S}}_{t,t-1}^T = \mathrm{cov}[\boldsymbol{Y}_t, \boldsymbol{X}_{t-1}|\boldsymbol{y}^{(1)}]$, $\widetilde{\boldsymbol{V}}_t^T = \mathrm{var}[\boldsymbol{X}_t|\boldsymbol{y}^{(1)}]$, $\widetilde{\boldsymbol{V}}_{t,t-1}^T = \mathrm{cov}[\boldsymbol{X}_t, \boldsymbol{X}_{t-1}|\boldsymbol{y}^{(1)}]$, and $\widetilde{\boldsymbol{V}}_{t-1,t}^T = \mathrm{cov}[\boldsymbol{X}_{t-1}, \boldsymbol{X}_t|\boldsymbol{y}^{(1)}]$. This gives the variance of both 'observed' model residuals (the ones associated with $\boldsymbol{y}^{(1)}$) and the unobserved model residuals (the ones associated with $\boldsymbol{y}^{(2)}$). When there are no missing values in $\boldsymbol{y}_t$, the $\widetilde{\mathbf{S}}_t^T$ and $\widetilde{\mathbf{S}}_{t,t-1}^T$ terms equal 0 and drop out.

[13] By extension, this is also the case for $\widetilde{\boldsymbol{V}}_t^T$, $\widetilde{\boldsymbol{V}}_{t,t-1}^T$, $\widetilde{\mathbf{S}}_t^T$ and $\widetilde{\mathbf{S}}_{t,t-1}^T$ which may seem counter-intuitive, but you can show it is true by working through the Kalman filter and smoother equations starting at t=1. Or run a Kalman filter/smoother algorithm with different data and the same parameters and you will see that the variances do not change with different data.

If the residuals are defined as in Harvey et al. (1998) with $\widehat{\boldsymbol{v}}_t$ on top and $\widehat{\boldsymbol{w}}_{t+1}$ on the bottom instead of $\widehat{\boldsymbol{w}}_t$, then

$$\widehat{\varepsilon}_t = \begin{bmatrix} \widehat{\boldsymbol{v}}_t \\ \widehat{\boldsymbol{w}}_{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_t - \mathbf{Z}_t \widetilde{\boldsymbol{x}}_t^T - \mathbf{a}_t \\ \widetilde{\boldsymbol{x}}_{t+1}^T - \mathbf{B}_{t+1} \widetilde{\boldsymbol{x}}_t^T - \mathbf{u}_{t+1} \end{bmatrix} \tag{55}$$

and the variance of $\widehat{\mathcal{E}}_t$ is

$$\left[\begin{array}{c|c} \mathrm{var}[\widehat{\boldsymbol{V}}_t] & \mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_{t+1}] \\ \hline (\mathrm{cov}[\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{W}}_{t+1}])^\top & \mathrm{var}[\widehat{\boldsymbol{W}}_{t+1}] \end{array} \right] \tag{56}$$

which is

$$\left[\begin{array}{c|c} \mathbf{R}_t - \mathbf{Z}_t \widetilde{\boldsymbol{V}}_t^T \mathbf{Z}_t^\top + \widetilde{\mathbf{S}}_t^T \mathbf{Z}_t^\top + \mathbf{Z}_t (\widetilde{\mathbf{S}}_t^T)^\top & -\widetilde{\mathbf{S}}_{t,t+1}^T + \widetilde{\mathbf{S}}_t^T \mathbf{B}_{t+1}^\top + \mathbf{Z}_t \widetilde{\boldsymbol{V}}_{t,t+1}^T - \mathbf{Z}_t \widetilde{\boldsymbol{V}}_t^T \mathbf{B}_{t+1}^\top \\ \hline (-\widetilde{\mathbf{S}}_{t,t+1}^T + \widetilde{\mathbf{S}}_t^T \mathbf{B}_{t+1}^\top + \mathbf{Z}_t \widetilde{\boldsymbol{V}}_{t,t+1}^T - \mathbf{Z}_t \widetilde{\boldsymbol{V}}_t^T \mathbf{B}_{t+1}^\top)^\top & \mathbf{Q}_{t+1} - \widetilde{\boldsymbol{V}}_{t+1}^T - \mathbf{B}_{t+1} \widetilde{\boldsymbol{V}}_t^T \mathbf{B}_{t+1}^\top + \widetilde{\boldsymbol{V}}_{t+1,t}^T \mathbf{B}_{t+1}^\top + \mathbf{B}_{t+1} \widetilde{\boldsymbol{V}}_{t,t+1}^T \end{array} \right]. \tag{57}$$

3 Harvey et al. 1998 algorithm for the conditional residuals

Harvey et al. (1998, pgs 112-113) give a recursive algorithm for computing the variance of the conditional residuals when the time-varying MARSS equation is written as:

$$\begin{gathered} \boldsymbol{x}_{t+1} = \mathbf{B}_{t+1} \boldsymbol{x}_t + \mathbf{u}_{t+1} + \mathbf{H}_{t+1} \epsilon_t, \\ \boldsymbol{y}_t = \mathbf{Z}_t \boldsymbol{x}_t + \mathbf{a}_t + \mathbf{G}_t \epsilon_t, \\ \text{where } \epsilon_t \sim \mathrm{MVN}(0, \mathbf{I}_{m+n \times m+n}) \\ \mathbf{H}_t \mathbf{H}_t^\top = \mathbf{Q}_t, \mathbf{G}_t \mathbf{G}_t^\top = \mathbf{R}_t, \text{ and } \mathbf{H}_t \mathbf{G}_t^\top = \mathrm{cov}[\boldsymbol{W}_t, \boldsymbol{V}_t] \end{gathered} \tag{58}$$

The $\mathbf{H}_t$ and $\mathbf{G}_t$ matrices specify the variance and covariance of $\boldsymbol{W}_t$ and $\boldsymbol{V}_t$. $\mathbf{H}_t$ has m rows and $m+n$ columns and $\mathbf{G}_t$ has n rows and $m+n$ columns. In the MARSS equation for this report (Equation 1), $\boldsymbol{W}_t$ and $\boldsymbol{V}_t$ are independent. To achieve this in the Harvey et al. form (Equation 58), the first n columns of $\mathbf{H}_t$ are all 0 and the last m columns of $\mathbf{G}_t$ are all zero.

The algorithm in Harvey et al. (1998) gives the variance of the 'normalized' residuals, the ϵ_t. I have modified their algorithm so it returns the 'non-normalized' residuals:

$$\varepsilon_t = \begin{bmatrix} \mathbf{G}_t \epsilon_t \\ \mathbf{H}_{t+1} \epsilon_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_t \\ \boldsymbol{w}_{t+1} \end{bmatrix}.$$

The Harvey et al. algorithm is a backwards recursion using the following output from the Kalman filter: the one-step ahead prediction covariance $\mathbf{F}_t$, the Kalman gain $\mathbf{K}_t$, $\widetilde{\boldsymbol{x}}_t^{t-1} = \mathrm{E}[\boldsymbol{X}_t | \boldsymbol{y}^{(1),1:t-1}]$ and $\widetilde{\boldsymbol{V}}_t^{t-1} = \mathrm{var}[\boldsymbol{X}_t | \boldsymbol{y}^{(1),1:t-1}]$. In the {MARSS} package, these are output from `MARSSkfss()` in `Sigma`, `Kt`, `xtt1` and `Vtt1`.

3.1 Algorithm for the non-normalized residuals

Start from $t = T$ and work backwards to $t = 1$. At time T, $r_T = 0_{1 \times m}$ and $N_T = 0_{m \times m}$. $\mathbf{B}_{t+1}$ and $\mathbf{Q}_{t+1}$ can be set to NA or 0. They will not appear in the algorithm at time T since $r_T = 0$ and $N_T = 0$. Note that the $\boldsymbol{w}$ residual and its associated variance and covariance with $\boldsymbol{v}$ at time T is NA since this residual would be for $\boldsymbol{x}_T$ to $\boldsymbol{x}_{T+1}$.

$$
\begin{aligned}
\mathbf{Q}'_{t+1} &= \begin{bmatrix} 0_{m\times n} & \mathbf{Q}_{t+1} \end{bmatrix},\ \mathbf{R}'_t = \begin{bmatrix} \mathbf{R}^*_t & 0_{n\times m} \end{bmatrix} \\
\mathbf{F}_t &= \mathbf{Z}^*_t \widetilde{\boldsymbol{V}}_t^{t-1} \mathbf{Z}_t^{*\top} + \mathbf{R}^*_t,\ n\times n \\
K_t = \mathbf{B}_{t+1}\mathbf{K}_t &= \mathbf{B}_{t+1} \widetilde{\boldsymbol{V}}_t^{t-1} \mathbf{Z}_t^{*\top} \mathbf{F}_t^{-1},\ m\times n \\
L_t &= \mathbf{B}_{t+1} - K_t \mathbf{Z}^*_t,\ m\times m \\
J_t &= \mathbf{Q}'_{t+1} - K_t \mathbf{R}'_t,\ m\times(n+m) \\
v_t &= \boldsymbol{y}^*_t - \mathbf{Z}_t \widetilde{\boldsymbol{x}}_t^{t-1} - \mathbf{a}_t,\ n\times 1 \\
u_t &= \mathbf{F}_t^{-1} v_t - K_t^\top r_t,\ n\times 1 \\
r_{t-1} &= \mathbf{Z}_t^{*\top} u_t + \mathbf{B}_{t+1}^\top r_t,\ m\times 1 \\
N_{t-1} &= \mathbf{Z}_t^{*\top} \mathbf{F}_t^{-1} \mathbf{Z}^*_t + L_t^\top N_t L_t,\ m\times m.
\end{aligned}
\tag{59}
$$

$\boldsymbol{y}^*_t$ is the observed data at time t with the i-th rows set to 0 if the i-th y is missing. Bolded terms are the same as in Equation 58 (and are output by `MARSSkfss()`). Unbolded terms are terms used in Harvey et al. (1998). The * on $\mathbf{Z}_t$ and $\mathbf{R}_t$, indicates that they are the missing value modified versions discussed in Shumway and Stoffer (2006, section 6.4) and Holmes (2012): to construct $\mathbf{Z}^*_t$ and $\mathbf{R}^*_t$, the rows of $\mathbf{Z}_t$ corresponding to missing rows of $\boldsymbol{y}_t$ are set to zero and the (i,j) and (j,i) terms of $\mathbf{R}_t$ corresponding the missing rows of $\boldsymbol{y}_t$ are set to zero. For the latter, this means if the i-th row of $\boldsymbol{y}_t$ is missing, then then the i-th row and column (including the value on the diagonal) in $\mathbf{R}_t$ are set to 0. Notice that $\mathbf{F}_t$ will have 0's on the diagonal if there are missing values. A modified inverse of $\mathbf{F}_t$ is used: any 0's on the diagonal of $\mathbf{F}_t$ are replaced with 1, the inverse is taken, and 1s on diagonals is replaced back with 0s.

The residuals (Harvey et al., 1998, eqn 24) are

$$
\widehat{\varepsilon}_t = \begin{bmatrix} \widehat{\boldsymbol{v}}_t \\ \widehat{\boldsymbol{w}}_{t+1} \end{bmatrix} = (\mathbf{R}'_t)^\top u_t + (\mathbf{Q}'_{t+1})^\top r_t \tag{60}
$$

The expected value of $\widehat{\mathcal{E}}_t$ is 0 and its variance is

$$
\widehat{\Sigma}]_t = \mathrm{var}[\widehat{\mathcal{E}}_t] = \mathbf{R}_t'^\top \mathbf{F}_t^{-1} \mathbf{R}'_t + J_t^\top N_t J_t. \tag{61}
$$

These $\widehat{\varepsilon}_t$ and $\widehat{\Sigma}]_t$ are for both the non-missing and missing $\boldsymbol{y}_t$. This is a modification to the Harvey et al. (1998) algorithm which does not give the variance for missing $\boldsymbol{y}$.

3.2 Difference in notation

In Equation 20 in Harvey et al. (1998), their T_t is my $\mathbf{B}_{t+1}$ and their $H_t H_t^\top$ is my $\mathbf{Q}_{t+1}$. Notice the difference in the time indexing. My time indexing on $\mathbf{B}$ and $\mathbf{Q}$ matches the left $\boldsymbol{x}$ while in theirs, T and H indexing matches the right $\boldsymbol{x}$. Thus in my implementation of their algorithm (Harvey et al., 1998, eqns. 21-24), $\mathbf{B}_{t+1}$ appears in place of T_t and $\mathbf{Q}_{t+1}$ appears in place of H_t. See comments below on normalization and the difference between $\mathbf{Q}$ and H.

Harvey et al. (1998, eqns. 19, 20) use G_t to refer to the $\mathrm{chol}(\mathbf{R}_t)^\top$ (non-zero part of the $n \times n+m$ matrix) and H_t to refer to $\mathrm{chol}(\mathbf{Q}_t)^\top$. I have replaced these with $\mathbf{R}'_t$ and $\mathbf{Q}'_t$ (Equation 59) which causes my variant of their algorithm (Equation 59) to give the 'non-normalized' variance of the residuals. The residuals function in the {MARSS} package has an option to give either normalized or non-normalized residuals.

$\mathbf{K}_t$ is the Kalman gain output by the {MARSS} package `MARSSkf()` function. The Kalman gain as used in the Harvey et al. (1998) algorithm is $K_t = \mathbf{B}_{t+1}\mathbf{K}_t$. Notice that Equation 21 in Harvey et al. (1998) has $H_t G_t^\top$ in the equation for K_t. This is the covariance of the state and observation errors, which is allowed to be non-zero given the way Harvey et al. write the errors in their Equations 19 and 20. The way the {MARSS} package model is written, the state and observation errors are independent of each other. Thus $H_t G_t^\top = 0$ and this term drops out of the K_t equation in Equation 59.

3.3 Computing the normalized residuals

To compute the normalized residuals and residuals variance, a block diagonal matrix with the inverse of the $\mathbf{R}$ and $\mathbf{Q}$ matrices is used. The normalized residuals are:

$$\begin{bmatrix} \mathbf{R}_t^{-1} & 0 \\ 0 & \mathbf{Q}_t^{-1} \end{bmatrix} \widehat{\varepsilon}_t \tag{62}$$

The variance of the normalized residuals is

$$\begin{bmatrix} \mathbf{R}_t^{-1} & 0 \\ 0 & \mathbf{Q}_t^{-1} \end{bmatrix} \widehat{\Sigma]}_t \begin{bmatrix} \mathbf{R}_t^{-1} & 0 \\ 0 & \mathbf{Q}_t^{-1} \end{bmatrix} \tag{63}$$

3.4 Computing the Cholesky standardized residuals

The Cholesky standardized residuals are computed by multiplying $\widehat{\varepsilon}_t$ by $(\widehat{\Sigma}_t)^{-1/2}$ (the inverse of the Cholesky decomposition of the variance-covariance matrix for $\widehat{\varepsilon}_t$):

$$\widehat{\varepsilon}_{t,std} = (\widehat{\Sigma}_t)^{-1/2}\widehat{\varepsilon}_t. \tag{64}$$

These residuals are uncorrelated (across the residuals at time t in $\widehat{\varepsilon}_t$). See Harvey and Koopman (1992) and Harvey et al. (1998, section V) for a discussion on how to use these residuals for outlier detection and structural break detection.

It is also common to use the marginal standardized residuals, which would be $\widehat{\varepsilon}_t$ multiplied by the inverse of the square-root of $\mathrm{dg}(\widehat{\Sigma}_t)$, where $\mathrm{dg}(\mathbf{A})$ is a diagonal matrix formed from the diagonal of the square matrix $\mathbf{A}$.

$$\widehat{\varepsilon}_{t,mar} = \mathrm{dg}(\widehat{\Sigma}_t)^{-1/2}\widehat{\varepsilon}_t \tag{65}$$

Marginal standardized residuals may be correlated at time t (unlike the Cholesky standardized residuals) but are a bit easier to interpret when there is correlation across the model residuals.

4 Distribution of the MARSS innovation residuals

One-step-ahead predictions (innovations) are often shown for MARSS models and these are used for likelihood calculations. Innovations are the difference between the data at time t minus the prediction of $\boldsymbol{y}_t$ given data up to $t-1$. This section gives the residual variance for the innovations and the analogous values for the states. Innovations residuals are the more common residuals discussed for state-space models; these are also known as the 'one-step-ahead' prediction residuals.

4.1 One-step-ahead model residuals

Define the innovations $\overline{\boldsymbol{v}}_t$ as:

$$\overline{\boldsymbol{v}}_t = E[\boldsymbol{Y}_t|\boldsymbol{y}_t^{(1)}] - \mathbf{Z}_t\widetilde{\boldsymbol{x}}_t^{t-1} - \mathbf{a}_t, \tag{66}$$

where $\widetilde{\boldsymbol{x}}_t^{t-1}$ is $\mathrm{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1),t-1}]$ (expected value of $\boldsymbol{X}_t$ conditioned on the data up to time $t-1$). The random variable, innovations over all possible $\boldsymbol{y}_t$, is $\overline{\boldsymbol{V}}_t$. Its mean is 0 and we want to find its variance.

This is conceptually different than the observed 'innovations'. First, this is the random variable 'innovation'. $\boldsymbol{y}_t^{(1)}$ here is not the actual data that you observe (the one data set that you have). It's the data that you could observe. $\boldsymbol{y}_t$ is a sample from the random variable $\boldsymbol{Y}_t$ and $\overline{\boldsymbol{v}}_t$ is a sample from the innovations you could observe. Second, $\overline{\boldsymbol{v}}_t$ includes both $\boldsymbol{y}_t^{(1)}$ and $\boldsymbol{y}_t^{(2)}$ (observed and unobserved $\boldsymbol{y}$). Normally, the innovations for missing data would appear as 0s, e.g., from a call to `MARSSfss()`. For missing data, $\overline{\boldsymbol{v}}_t$ is not necessarily 0. For example if $\boldsymbol{y}$ is multivariate and correlated with each other through $\mathbf{R}$ or a shared $\boldsymbol{x}$ dependency.

The derivation of the variance of $\overline{\boldsymbol{V}}_t$ follows the exact same steps as the smoothations $\widehat{\boldsymbol{V}}_t$, except that we condition on the data up to $t-1$ not up to T. Thus using Equation 25, we can write the variance directly as:

$$\mathrm{var}[\overline{\boldsymbol{V}}_t] = \mathbf{R}_t - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^{t-1}\mathbf{Z}_t^\top + \widetilde{\mathbf{S}}_t^{t-1}\mathbf{Z}_t^\top + \mathbf{Z}_t(\widetilde{\mathbf{S}}_t^{t-1})^\top \tag{67}$$

where the $\widetilde{\boldsymbol{V}}_t^{t-1}$ and $\widetilde{\mathbf{S}}_t^{t-1}$ are now conditioned on only the data from 1 to $t-1$. $\widetilde{\mathbf{S}}_t^{t-1} = \mathrm{cov}[\boldsymbol{Y}_t, \boldsymbol{X}_t|\boldsymbol{y}^{(1),t-1}] = \mathrm{cov}[\mathbf{Z}_t\boldsymbol{X}_t + \mathbf{a}_t + \boldsymbol{V}_t, \boldsymbol{X}_t|\boldsymbol{y}^{(1),t-1}]$. $\boldsymbol{y}_t$ is not in the conditional since it only includes data up to $t-1$. Without

$\boldsymbol{y}_t$ in the conditional, $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ and by extension $\boldsymbol{V}_t$ and $\boldsymbol{X}_t$ are independent[14] and $\text{cov}[\mathbf{Z}_t\boldsymbol{X}_t + \mathbf{a}_t + \boldsymbol{V}_t, \boldsymbol{X}_t|\boldsymbol{y}^{(1),t-1}] = \text{cov}[\mathbf{Z}_t\boldsymbol{X}_t, \boldsymbol{X}_t|\boldsymbol{y}^{(1),t-1}] = \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^{t-1}$. Therefore, $\mathbf{Z}_t(\widetilde{\mathbf{S}}_t^{t-1})^\top = \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^{t-1}\mathbf{Z}_t^\top = \widetilde{\mathbf{S}}_t^{t-1}(\mathbf{Z}_t)^\top$. Thus Equation 67 reduces to

$$\text{var}[\overline{\boldsymbol{V}}_t] = \mathbf{R}_t + \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^{t-1}\mathbf{Z}_t^\top. \tag{68}$$

Note $\text{var}[\overline{\boldsymbol{V}}_t]$ is a standard output from Kalman filter functions and is used to compute the likelihood of the data (conditioned on a set of parameters). In the Kalman filter in the {MARSS} package, it is output as `Sigma` from `MARSSkfss()`.

4.2 One-step ahead state residuals

Define the state residuals conditioned on the data from 1 to $t-1$ as $\overline{\boldsymbol{w}}_t$.

$$\overline{\boldsymbol{w}}_t = \widetilde{\boldsymbol{x}}_t^t - \mathbf{B}_t\widetilde{\boldsymbol{x}}_{t-1}^{t-1} - \mathbf{u}_t, \tag{69}$$

where $\widetilde{\boldsymbol{x}}_{t-1}^{t-1}$ is $\text{E}[\boldsymbol{X}_{t-1}|\boldsymbol{y}^{(1),t-1}]$ (expected value of $\boldsymbol{X}_{t-1}$ conditioned on the data up to time $t-1$) and $\widetilde{\boldsymbol{x}}_t^t$ is $\text{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1),t}]$ (expected value of $\boldsymbol{X}_t$ conditioned on the data up to time t). From the Kalman filter equations:

$$\widetilde{\boldsymbol{x}}_t^t = \mathbf{B}_t\widetilde{\boldsymbol{x}}_{t-1}^{t-1} + \mathbf{u}_t + \mathbf{K}_t\overline{\boldsymbol{v}}_t \tag{70}$$

Thus, $\overline{\boldsymbol{w}}_t$ is a transformed $\overline{\boldsymbol{v}}_t$:

$$\overline{\boldsymbol{w}}_t = \mathbf{K}_t\overline{\boldsymbol{v}}_t, \tag{71}$$

where $\mathbf{K}_t$ is the Kalman gain. $\mathbf{K}_t = \widetilde{\boldsymbol{V}}_t^{t-1}\mathbf{Z}_t^\top[\mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^{t-1}\mathbf{Z}_t^\top + \mathbf{R}_t]^{-1}$ and $\mathbf{Z}_t$ is the missing values modified $\mathbf{Z}_t$ where if the i-th $\boldsymbol{y}_t$ is missing, the i-th row of $\mathbf{Z}_t$ is set to all 0s (Shumway and Stoffer 2006, equation 6.78). Thus the variance of $\overline{\boldsymbol{W}}_t$ is

$$\text{var}[\overline{\boldsymbol{W}}_t] = \mathbf{K}_t\,\text{var}[\overline{\boldsymbol{V}}_t]\mathbf{K}_t^\top \tag{72}$$

4.3 Joint distribution of the conditional one-step-ahead residuals

4.3.1 with the state residuals defined from $t-1$ to t

Define the one-step ahead residuals as

$$\overline{\varepsilon}_t = \begin{bmatrix}\overline{\boldsymbol{v}}_t \\ \overline{\boldsymbol{w}}_t\end{bmatrix} \tag{73}$$

The covariance of $\overline{\boldsymbol{V}}_t$ and $\overline{\boldsymbol{W}}_t$ is

$$\text{cov}[\overline{\boldsymbol{V}}_t, \overline{\boldsymbol{W}}_{t+1}] = \text{cov}[\overline{\boldsymbol{V}}_t, \mathbf{K}_t\overline{\boldsymbol{V}}_t] = \text{var}[\overline{\boldsymbol{V}}_t]\mathbf{K}_t^\top \tag{74}$$

The joint variance-covariance matrix is

$$\overline{\Sigma}_t = \text{var}[\overline{\varepsilon}_t] = \left[\begin{array}{c|c} \text{var}[\overline{\boldsymbol{V}}_t] & \text{var}[\overline{\boldsymbol{V}}_t]\mathbf{K}_t^\top \\ \hline \mathbf{K}_t\,\text{var}[\overline{\boldsymbol{V}}_t] & \mathbf{K}_t\,\text{var}[\overline{\boldsymbol{V}}_t]\mathbf{K}_t^\top \end{array}\right], \tag{75}$$

Since $\overline{\boldsymbol{w}}_t = \mathbf{K}_t\overline{\boldsymbol{v}}_t$, the state one-step-ahead residuals are perfectly correlated with the model one-step-ahead residuals so the joint distribution is not useful (all the information is in the variance of $\overline{\boldsymbol{V}}_t$).

The Cholesky standardized residuals for $\overline{\varepsilon}_t$ are

$$\overline{\Sigma}_t^{-1/2}\overline{\varepsilon}_t \tag{76}$$

However the Cholesky standardized joint residuals cannot be computed since $\overline{\Sigma}_t$ is not positive definite. Because $\overline{\boldsymbol{w}}_t$ equals $\mathbf{K}_t\overline{\boldsymbol{v}}_t$, the state residuals are completely explained by the model residuals. However we can compute the Cholesky standardized model residuals using

$$\text{var}[\overline{\boldsymbol{V}}_t]^{-1/2}\overline{\boldsymbol{v}}_t \tag{77}$$

[14] This is only true given the way the MARSS equation is written in this report where V_t and W_t are independent. This is not the case for the more general Harvey et al. MARSS model which allows covariance between V_t and W_t.

$\text{var}[\overline{\boldsymbol{V}}_t]^{-1/2}$ is the inverse of the lower triangle of the Cholesky decomposition of $\text{var}[\overline{\boldsymbol{V}}_t]$.

The marginal standardized joint residuals for $\overline{\varepsilon}_t$ (both model and state one-step ahead residuals) can be computed with the inverse of the square root of the diagonal of $\overline{\Sigma}_t$:

$$\text{dg}(\overline{\Sigma}_t)^{-1/2}\overline{\varepsilon}_t \tag{78}$$

where $dg(\mathbf{A})$ is a diagonal matrix formed from the diagonal of the square matrix $\mathbf{A}$.

4.3.2 with the state residuals defined from t to $t+1$

Note that the model residual is conditioned on data 1 to $t-1$ and the state residual on data 1 to t in this case.

Define the one-step ahead residuals as

$$\overline{\varepsilon}_t^* = \begin{bmatrix} \overline{\boldsymbol{v}}_t \\ \overline{\boldsymbol{w}}_{t+1} \end{bmatrix} \tag{79}$$

The covariance of $\overline{\boldsymbol{V}}_t$ and $\overline{\boldsymbol{W}}_{t+1}$ is

$$\text{cov}[\overline{\boldsymbol{V}}_t, \overline{\boldsymbol{W}}_{t+1}] = \text{cov}[\overline{\boldsymbol{V}}_t, \mathbf{K}_{t+1}\overline{\boldsymbol{V}}_{t+1}] = \text{cov}[\overline{\boldsymbol{V}}_t, \overline{\boldsymbol{V}}_{t+1}]\mathbf{K}_{t+1}^\top \tag{80}$$

Innovations residuals are temporally uncorrelated, thus $\text{cov}[\overline{\boldsymbol{V}}_t, \overline{\boldsymbol{V}}_{t+1}] = 0$ and thus $\text{cov}[\overline{\boldsymbol{V}}_t, \overline{\boldsymbol{W}}_{t+1}] = 0$.

The joint variance-covariance matrix is

$$\overline{\Sigma}_t^* = \text{var}[\overline{\varepsilon}_t^*] = \left[\begin{array}{c|c} \text{var}[\overline{\boldsymbol{V}}_t] & 0 \\ \hline 0 & \mathbf{K}_{t+1}\,\text{var}[\overline{\boldsymbol{V}}_{t+1}]\mathbf{K}_{t+1}^\top \end{array}\right], \tag{81}$$

The Cholesky standardized residuals for $\overline{\varepsilon}_t^*$ are

$$(\overline{\Sigma}_t^*)^{-1/2}\overline{\varepsilon}_t^* \tag{82}$$

$\overline{\Sigma}_t^*$ is positive definite so its Cholesky decomposition can be computed.

The marginal standardized joint residuals for $\overline{\varepsilon}_t^*$ are:

$$\text{dg}(\overline{\Sigma}_t^*)^{-1/2}\overline{\varepsilon}_t^* \tag{83}$$

where $dg(\mathbf{A})$ is a diagonal matrix formed from the diagonal of the square matrix $\mathbf{A}$.

5 Distribution of the MARSS contemporaneous model residuals

Contemporaneous model residuals are the difference between the data at time t minus the prediction of $\boldsymbol{y}_t$ given data up to t. This section gives the residual variance for these residuals. There are no state residuals for this case as that would require the expected value of $\boldsymbol{X}_t$ conditioned on the data up to $t+1$.

Define the contemporaneous model residuals $\dot{\boldsymbol{v}}_t$ as:

$$\dot{\boldsymbol{v}}_t = E[\boldsymbol{Y}_t|\boldsymbol{y}_t^{(1)}] - \mathbf{Z}_t\widetilde{\boldsymbol{x}}_t^t - \mathbf{a}_t, \tag{84}$$

where $\widetilde{\boldsymbol{x}}_t^t$ is $\text{E}[\boldsymbol{X}_t|\boldsymbol{y}^{(1),t}]$ (expected value of $\boldsymbol{X}_t$ conditioned on the data up to time t). The random variable, contemporaneous model residuals over all possible $\boldsymbol{y}_t$, is $\dot{\boldsymbol{V}}_t$. Its mean is 0 and we want to find its variance.

The derivation of the variance of $\dot{\boldsymbol{V}}_t$ follows the exact same steps as the smoothations $\widehat{\boldsymbol{V}}_t$, except that we condition on the data up to t not up to T. Thus using Equation 25, we can write the variance directly as:

$$\text{var}[\dot{\boldsymbol{V}}_t] = \mathbf{R}_t - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^t\mathbf{Z}_t^\top + \widetilde{\mathbf{S}}_t^t\mathbf{Z}_t^\top + \mathbf{Z}_t(\widetilde{\mathbf{S}}_t^t)^\top \tag{85}$$

where the $\widetilde{\boldsymbol{V}}_t^t$ and $\widetilde{\mathbf{S}}_t^t$ are now conditioned on only the data from 1 to t. $\widetilde{\mathbf{S}}_t^t = \text{cov}[\boldsymbol{Y}_t, \boldsymbol{X}_t|\boldsymbol{y}^{(1),t}] = \text{cov}[\mathbf{Z}_t\boldsymbol{X}_t + \mathbf{a}_t + \boldsymbol{V}_t, \boldsymbol{X}_t|\boldsymbol{y}^{(1),t}]$. If $\boldsymbol{y}_t$ has no missing values, this reduces to $\mathbf{R}_t - \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^t\mathbf{Z}_t^\top$ while if $\boldsymbol{y}_t$ is all missing values, this reduces to $\mathbf{R}_t + \mathbf{Z}_t\widetilde{\boldsymbol{V}}_t^t\mathbf{Z}_t^\top$. See discussion of this after Equation 25.

Equation 85 gives the equation for the case where $\boldsymbol{y}_t$ is partially observed. $\widetilde{\mathbf{S}}_t^t$ is output by the `MARSShatyt()` function in the {MARSS} package and $\widetilde{\boldsymbol{V}}_t^t$ is output by the `MARSSkfss()` function.

References

Commandeur, J. J. and Koopman, S. J. (2007). *An introduction to state space time series analysis*. Practical Econometrics. Oxford University Press, Oxford.

de Jong, P. (1988). A cross-validation filter for time series models. *Biometrika*, 75(3):594–600.

de Jong, P. and Penzer, J. (1998). Diagnosing shocks in time series. *Journal of the American Statistical Association*, 93(442):796–806.

Harvey, A. and Koopman, S. J. (1992). Diagnostic checking of unobserved-components time series models. *Journal of Business and Economic Statistics,*, 10:377–389.

Harvey, A., Koopman, S. J., and Penzer, J. (1998). Messy time series: a unified approach. *Advances in Econometrics*, 13:103–143.

Holmes, E. E. (2012). Derivation of the EM algorithm for constrained and unconstrained MARSS models. Technical report, arXiv:1302.3919 [stat.ME].

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