Deka, B., and Goulet, J-A. (2023).*Approximate Gaussian Variance Inference for State-Space Models. International Journal of Adaptive Control and Signal Processing. 1-29 doi*

Approximate Gaussian Variance Inference for State-Space Models

Bhargob Deka* and James-A. Goulet
Department of Civil, Geologic and Mining Engineering
Polytechnique Montreal, CANADA

August 11, 2023

Abstract

State-space models require an accurate knowledge of the process error ($\mathbf{Q}$) and measurement error ($\mathbf{R}$) covariance matrices for exact state estimation. Even though the matrix $\mathbf{R}$ can be, in many situations, considered to be known from the measuring instrument specifications, it is still a challenge to infer the $\mathbf{Q}$ matrix online while providing reliable estimates along with a low computational cost. In this paper, we propose an analytically tractable online Bayesian inference method for inferring the $\mathbf{Q}$ matrix in state-space models. We refer to this method as *approximate Gaussian variance inference* (AGVI) using which we are able to treat the error variance and covariance terms in the full $\mathbf{Q}$ matrix as Gaussian hidden states and infer them simultaneously with the other hidden states in a closed-form manner. The two case studies show that the method is able to provide statistically consistent estimates for the mean and uncertainties of the error variance terms for univariate and multivariate cases. The method also exceeds the performance of the existing adaptive Kalman filter methods both in terms of accuracy and computational efficiency.

Keywords: state-space models, Bayesian inference, Gaussian multiplicative approximation, closed-form inference, process error covariance matrix, online parameter estimation.

1 Introduction

For linear dynamic systems, the Kalman filter is an exact state estimator if the process error ($\mathbf{Q}$) and the measurement error ($\mathbf{R}$) covariance matrices are known [30]. In most practical situations, the deterministic part of the model which includes the transition and the observation models is formulated based on known system dynamics. In contrast, the stochastic part representing the $\mathbf{Q}$ and $\mathbf{R}$ matrices is either unknown or only approximately known [14, 29, 30]. Previous studies have also shown that using incorrect error covariance matrices may result in large estimation errors or even cause divergence [30, 32, 46]. Even though in many situations, the matrix $\mathbf{R}$ can be considered to be known from measuring instrument specification, the $\mathbf{Q}$ matrix is often unknown. Hence, an accurate estimation of the matrix $\mathbf{Q}$ is necessary for the exact state estimation [31, 32].

This paper provides an analytical Bayesian inference method called the *approximate Gaussian variance inference* (AGVI) for performing closed-form online estimation of the error variance and covariance terms in the full $\mathbf{Q}$ matrix. By definition, the expected value of the square of the univariate process error W^2 is equal to the error variance parameter, i.e., $\mathbb{E}[W^2] = \sigma_W^2$, given that W has a zero mean. With the approximation that W^2 is Gaussian such that $W^2 \sim \mathcal{N}(w^2; \mathbb{E}[W^2], \text{var}(W^2))$,

*Corresponding author: bhargobdeka11@gmail.com

the error variance parameter is the same as the mean parameter for the probability density function (PDF) of W^2. Subsequently, considering that this mean parameter $\mathbb{E}[W^2]$ is a random variable itself, inferring its posterior becomes analogous to computing the posterior for the error variance term. The AGVI method utilizes this definition and formulates the relationship between the process error W, the square of process error W^2, and $\mathbb{E}[W^2]$ by leveraging the Gaussian multiplicative approximation (GMA) [9, 18] that provides the exact moments for W^2. Thereafter, the Gaussian conjugate prior [16, 33] is used to analytically infer the unknown mean parameter for W^2, i.e., $\mathbb{E}[W^2] = \sigma_W^2$, using closed-form equations. The methodology is also extended to the multivariate observation model case where one error variance term σ_W^2 is inferred for each observation equation along with the covariance between each pair of process error terms. The paper also provides a closed-form square-root filtering technique using the Cholesky decomposition such that the estimated $\mathbf{Q}$ matrix is always positive semi-definite (PSD).

The layout of this paper is as follows: Section 2 presents related works; Section 3 provides the AGVI method for estimating the univariate process error variance, and Section 4 extends the methodology to the multivariate case for inferring the full $\mathbf{Q}$ matrix. Section 5 presents two case studies showing the application of the method; Section 5.1 presents the online estimation for the univariate error variance term and Section 5.2 shows the applicability of AGVI in inferring the full $\mathbf{Q}$ matrix for multiple time series.

1.1 Notations

The following notations are used throughout the manuscript; x : lowercase denotes a variable, X : slanted uppercase denotes a random variable, $\mathbf{X}$: bold upright uppercase denotes a deterministic matrix, $\boldsymbol{x}$: bold lowercase denotes a vector, $\boldsymbol{X}$: bold slanted uppercase denotes vector of random variables, $\boldsymbol{y}_{1:t}$: denote observations from 1 to t, $\boldsymbol{\mu}$: mean vector, $\boldsymbol{\Sigma}$: covariance matrix, W : random process error, W^2: square of process error W, $\overline{W^2}$: random variable representing the expected value for W^2, W^iW^j: product of any i^{th} and j^{th} process error, $\overline{W^iW^j}$: random variable for the expected value of W^iW^j, $\boldsymbol{W^p}$: random vector of all the product terms W^iW^j, $\overline{\boldsymbol{W^p}}$: random vector of all expected values of $\boldsymbol{W^p}$, $\boldsymbol{\Sigma^{XW}}$: covariance matrix between $\boldsymbol{X}$ and $\boldsymbol{W}$, $\boldsymbol{L^W}$: upper triangular random matrix, $\overrightarrow{\boldsymbol{L^W}}$: random vector of all elements in $\boldsymbol{L^W}$.
Throughout the article, consistency is maintained in the notations as we transition from the univariate to the multivariate case study. For instance, we use W^2 to represent the random variable denoting the square of the univariate random process error W. Similarly, W^iW^j represents the random variable for the product of W^i and W^j, where the superscripts indicate the i^{th} and j^{th} error terms. To denote the random expected value for W^iW^j, we use a bar on top, denoted by $\overline{W^iW^j}$. In the case of the univariate scenario, this becomes $\overline{W^2}$, while for any i^{th} process error, it is expressed as $\overline{(W^i)^2}$. The expected value and variance for $\overline{W^iW^j}$ are denoted as $\mu^{\overline{W^iW^j}}$ and $(\sigma^{\overline{W^iW^j}})^2$, respectively, with the corresponding random variable provided in the superscript.

2 Related Works

The adaptive Kalman filters (AKF) were developed to estimate both the states and the error covariance matrices together by adaptively adjusting the Kalman filter to the measured data such that the estimation errors can be either bounded or reduced [32]. The AKFs are broadly grouped as follows: 1) *correlation methods* [5, 6, 13, 30, 34, 37, 47], 2) *covariance-matching methods* (CMM) [2, 35, 39], 3) *maximum likelihood methods* [23, 43], and 4) *Bayesian methods* [21, 22, 29, 38, 42]. One such AKF is the *innovation correlation method* (ICM) [6, 30] that uses the auto-correlation

function of the innovations to form a system of linear equations involving the unknown covariance matrices. A least-square method is then used to solve these equations simultaneously to obtain the estimates for the $\mathbf{Q}$ and $\mathbf{R}$ matrices. The correlation methods provide unbiased estimates for linear time-invariant (LTI) systems and only asymtotically unbiased estimates for linear time-varying (LTV) systems [10]. A recent correlation approach called the *measurement difference method* (MDM) [10, 11, 13, 14, 26] was proposed that is capable of providing unbiased and weakly consistent estimates for LTV systems, even for datasets with small number of measurements [10, 26]. MDM does not require any prior knowledge on the error covariance matrices and can also be performed online using recursive least-square methods [24, 26]. On the other hand, there are CMM methods such as the *adaptive limited memory filter* (ALMF) [35] that computes sample covariance matrices at each time step for both the state prediction error and the innovation sequence using either the entire past data or over a moving window. However, such methods have shown to produce biased estimates for the covariance matrices and often fail to ensure the positive-definiteness of matrices when the sample size of the data is small [4, 37]. Shumway and Stoffer [43] provided a framework that uses the *expectation-maximization* (EM) algorithm [40] to obtain both the states and the error covariance matrices even when the data is irregularly spaced, but this can only be applied offline [4]. An extensive amount of literature exists for the AKF methods under the Bayesian category. The Bayesian methods include *state augmentation methods* primarily relying on nonlinear estimation techniques such as the *extended Kalman filter* (EKF) [4], the *unscented Kalman filter* (UKF) [25], or the *particle filters* [36,45] for the joint estimation of both the states and the error covariance matrices (ECM). While most methods in this category identify the error variances offline, Kontoroupi and Smyth [25] provided an online estimation method by employing an approximation of the inverse gamma conjugacy. Online Bayesian inference methods also exists that relies on Markov chain Monte Carlo moves within a particle filter but they suffer from the particle degeneracy problem [15, 22, 45]. Another Bayesian method is the *interactive multiple models* (IMM) [27] that defines multiple models each having a separate dynamic model with its own ECM as well as the transitional probabilities between one model i and another model j at any given time step t. A set of several Kalman filters are run on parallel to evaluate the state estimates for each model simultaneously. However, exact estimates using IMM can be obtained only when an infinitely large numbers of models are considered [42].

The *variational Bayes* (VB) methods have been proposed to approximate the intractable joint posterior PDF of the states and the covariance matrices at a comparatively lower computational cost than using the particle filters or the multiple model methods [14, 42]. Sarkaa and Hartikainen [41] proposed the VB-AKF method to obtain the full $\mathbf{R}$ matrix using an inverse Wishart conjugate prior. However, the same methodology could not be applied to obtain the $\mathbf{Q}$ matrix, since it does not appear in simple conjugate prior form, as opposed to the $\mathbf{R}$ matrix [28, 41]. Ardeshiri et al. [1] proposed a VB based RTS smoother to obtain both the $\mathbf{Q}$ and $\mathbf{R}$ matrices, but it can only evaluate the error covariance matrices offline [19, 28]. Huang et al. [19] proposed an online VB-AKF method, referred to as VBAKF-PR, to directly estimate the joint distribution of the states, the state prediction error covariance matrix, and the $\mathbf{R}$ matrix by using the conjugacy of the inverse Wishart prior for the covariance matrices. However, the method requires an accurate nominal $\mathbf{Q}$ matrix based on problem-specific expertise without which the performance degrades drastically. Moreover, the method has additional parameters such as the tuning parameter, the forgetting factor, and the number of fixed-point iterations per time step that needs to be tuned. The *sliding window variational adaptive Kalman filter* (SWVAKF) method overcomes the limitation of VBAKF-PR as it is robust to the initialization of the nominal $\mathbf{Q}$ matrix and proved to be computationally more efficient by avoiding the fixed-point iteration step. Hence, there are several methods that have been proposed to estimate the error covariance matrices. However, most methods

are either offline in nature [30,43], restricted to linear dynamic systems [19,21] or are computationally demanding [27, 40, 45]. Furthermore, there is no closed-form method to obtain these matrices and none of the available methods have demonstrated the capacity to estimate a high-dimensional full $\mathbf{Q}$ matrix. Hence, there is still the challenge to develop a method that performs closed-form online estimation of the matrix $\mathbf{Q}$ and that is scalable to high-dimensional domains.

3 Univariate Process Error

This section presents the mathematical formulation of the AGVI method for inferring the variance parameter σ_W^2 associated with the univariate process error $W \sim \mathcal{N}(w; 0, \sigma_W^2)$ in the context of state-space models.

3.1 Problem Formulation

Let us consider an $\mathtt{N}$-dimensional hidden state vector at time $t-1$, $\boldsymbol{x}_{t-1} = [x_1\ x_2\ \cdots x_{\mathtt{N}}]_{t-1}^\intercal$, having a Gaussian PDF such that $\boldsymbol{X}_{t-1|t-1} \sim \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{t-1|t-1}, \boldsymbol{\Sigma}_{t-1|t-1})$ where $\boldsymbol{\mu}_{t-1|t-1} = \mathbb{E}[\boldsymbol{X}_{t-1}|\boldsymbol{y}_{1:t-1}]$ is the prior mean and $\boldsymbol{\Sigma}_{t-1|t-1} = \text{var}(\boldsymbol{X}_{t-1}|\boldsymbol{y}_{1:t-1})$ is the prior covariance matrix. Note that for brevity, the notation $\boldsymbol{X}_{t-1|t-1}$ is used as a shorthand for $\boldsymbol{X}_{t-1}|\boldsymbol{y}_{1:t-1}$. The transition and the observation equations for the linear Gaussian state-space models [17], are given by

$$\begin{array}{rclcrcl} \boldsymbol{x}_t & = & \mathbf{A}\boldsymbol{x}_{t-1} + \boldsymbol{w}_t, & \boldsymbol{w} & : & \boldsymbol{W} & \sim & \mathcal{N}(w; 0, \mathbf{Q}), \\ y_t & = & \mathbf{C}\boldsymbol{x}_t + v_t, & v & : & V & \sim & \mathcal{N}(v; 0, \mathbf{R}), \end{array} \tag{1}$$

where $\mathbf{A}$ is the transition matrix, $\boldsymbol{w}_t = [w_1\ w_2\ \cdots w_{\mathtt{N}}]_t^\intercal$ is a vector of process error terms for which $\mathbf{Q}$ is the process error covariance matrix, y_t is the observation, $\mathbf{C}$ is the observation matrix, and v_t is the observation error for which the observation error variance is $\mathbf{R} = \sigma_V^2$. The $\mathbf{A}$ and $\mathbf{Q}$ matrices are constructed by assembling $\mathtt{S}$ specific components given by

$$\begin{aligned} \mathbf{A} &= \text{blkdiag}(\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_{\mathtt{S}}), \\ \mathbf{Q} &= \text{blkdiag}(\mathbf{Q}_1, \mathbf{Q}_2, \cdots, \mathbf{Q}_{\mathtt{S}}), \end{aligned} \tag{2}$$

where blkdiag$(\cdot,\cdot)$ refers to block diagonal assembly of the individual components. The $\mathbf{Q}$ matrix in Equation 2 can be further described by

$$\mathbf{Q} = \text{blkdiag}(\mathbf{Q}_1(\sigma_{W_1}^2, \Delta t), \mathbf{Q}_2(\sigma_{W_2}^2, \Delta t), \cdots, \mathbf{Q}_{\mathtt{S}}(\sigma_{W_{\mathtt{S}}}^2, \Delta t)), \tag{3}$$

where each component $\mathbf{Q}_i(\sigma_{W_i}^2, \Delta t)$ can be represented as a function of the error variance parameter $\sigma_{W_i}^2$ and the time difference between two successive observations Δt. For example, consider that the model comprises two generic components, namely the local trend ($\mathtt{LT}$) and the autoregressive ($\mathtt{AR}$), for which the global $\mathbf{Q}$ matrix is

$$\mathbf{Q} = \begin{bmatrix} \sigma_{\mathtt{LT}}^2 \cdot \begin{bmatrix} \frac{\Delta t^4}{4} & \frac{\Delta t^3}{2} \\ \frac{\Delta t^3}{2} & \Delta t^2 \end{bmatrix} & \mathbf{0} \\ \mathbf{0} & \sigma_{\mathtt{AR}}^2 \end{bmatrix}, \tag{4}$$

where the $\mathbf{Q}$ matrices associated with the local trend $\mathbf{Q}_1$ and the autoregressive component $\mathbf{Q}_2$ are

$$\mathbf{Q}_1 = \sigma_{\mathtt{LT}}^2 \cdot \begin{bmatrix} \frac{\Delta t^4}{4} & \frac{\Delta t^3}{2} \\ \frac{\Delta t^3}{2} & \Delta t^2 \end{bmatrix}, \quad \mathbf{Q}_2 = \sigma_{\mathtt{AR}}^2,$$

where $\sigma^2_{\texttt{LT}}$ and $\sigma^2_{\texttt{AR}}$ are the process error variance terms associated with the `LT` and the `AR` components, respectively [17]. Both matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are assembled in a block diagonal arrangement to get the $\mathbf{Q}$ matrix as shown by Equation 4. Moreover, for a time series consisting of a single observation variable y_t, it is only possible to infer σ^2_W for one component $\mathbf{Q}_i$, while all other should be either known or 0. This is because only a single unknown variable can be uniquely solved per equation. Hence, for each time series there is one unique process error variance that can be inferred. The next section describes the various steps for performing AGVI in order to obtain the posterior PDF for a single error variance parameter.

3.2 Methodology

The proposed method considers the expected value of W^2 as a Gaussian random variable represented by $\overline{W^2}$ such that

$$\overline{W^2} \sim \mathcal{N}\big(\overline{w^2}; \mu^{\overline{W^2}}, (\sigma^{\overline{W^2}})^2\big), \tag{5}$$

where $\mu^{\overline{W^2}}$ and $(\sigma^{\overline{W^2}})^2$ are the hyperprior mean and variance for $\overline{W^2}$. Using Equation 5, the PDF of W can be re-written as

$$W \sim \mathcal{N}(w; 0, \overline{w^2}). \tag{6}$$

Hence, the first objective is to obtain the marginal PDF of W such that the random variance $\overline{W^2}$ can be marginalized out. The following lemmas are invoked to show that the marginal PDF of W can be obtained using the marginal PDF of W^2. The subsequent proposition uses these lemmas to provide the prior predictive PDF for W at a time t such that $W_{t|t-1} \sim \mathcal{N}(w_t; 0, \mu^{\overline{W^2}}_{t-1|t-1})$.

Lemma 1. *Given that W is Gaussian with a zero mean and W^2 is approximated as a Gaussian random variable given by $W^2 \sim \mathcal{N}(w^2; \mu^{W^2}, (\sigma^{W^2})^2)$ for which the exact moments are provided by GMA (see Appendix A), it can be shown that the PDF of W^2 is dependent only on the mean parameter μ^{W^2} so that*

$$W^2 \sim \mathcal{N}\big(w^2; \mu^{W^2}, 2(\mu^{W^2})^2\big),$$

where the variance term $(\sigma^{W^2})^2$ is equal to $2(\mu^{W^2})^2$. As a result, the PDF $f(w^2|\mu^{W^2}, (\sigma^{W^2})^2)$ can be shown by $f(w^2|\mu^{W^2})$.

Proof. See Appendix B.1. □

Lemma 2. *Given that the parameter μ^{W^2} in $f(w^2|\mu^{W^2})$ is considered as a Gaussian random variable $\overline{W^2} \sim \mathcal{N}(\overline{w^2}; \mu^{\overline{W^2}}, (\sigma^{\overline{W^2}})^2)$, the mean and variance of the prior predictive PDF of $W^2_{t|t-1}$ are given by*

$$\begin{aligned} \mu^{W^2}_{t|t-1} &= \mu^{\overline{W^2}}_{t-1|t-1}, \\ (\sigma^{W^2}_{t|t-1})^2 &= 3(\sigma^{\overline{W^2}}_{t-1|t-1})^2 + 2(\mu^{\overline{W^2}}_{t-1|t-1})^2, \end{aligned}$$

where $\mu^{\overline{W^2}}_{t-1|t-1}$ and $(\sigma^{\overline{W^2}}_{t-1|t-1})^2$ are the prior moments for $\overline{W^2}_{t-1|t-1}$.

Proof. See Appendix B.2. □

Proposition 1. *Considering that the mean parameter μ^{W^2} is itself a random variable $\overline{W^2}$ so that*

$$\overline{W^2_{t-1|t-1}} \sim \mathcal{N}\big(\overline{w^2_{t-1}}\,; \mu^{\overline{W^2}}_{t-1|t-1}, (\sigma^{\overline{W^2}}_{t-1|t-1})^2\big),$$

where $\mu^{\overline{W^2}}_{t-1|t-1}$ and $(\sigma^{\overline{W^2}}_{t-1|t-1})^2$ are the hyperprior mean and variance for $\overline{W^2}_{t-1|t-1}$, the error variance σ^2_W can be made equal to

$$\sigma^2_W = \mu^{\overline{W^2}}_{t-1|t-1}. \tag{7}$$

Proof. Using Lemmas 1 & 2, and considering the one-to-one relationship between the moments of W and W^2, the prior predictive PDF of $W_{t|t-1}$ can be formulated as

$$f(w_t) = \mathcal{N}(w_t; 0, \mu^{\overline{W^2}}_{t-1|t-1}),$$

where by Lemma 2, the variance of $W_{t|t-1}$ is $\sigma^2_W = \mathbb{E}[W^2_{t|t-1}] = \mu^{\overline{W^2}}_{t-1|t-1}$. □

The next objective is to perform the prediction step in the filtering procedure using the model matrices $\mathbf{A}$, $\mathbf{C}$, $\mathbf{Q}$, and $\mathbf{R}$ defined in Section 3.1 and the prior knowledge of σ^2_W. The transition model for $\overline{w^2}$ is $\overline{w^2}_t = \overline{w^2}_{t-1}$, where the hidden state $\overline{w^2}$ is assumed to be constant from $t-1$ to t. Using the prior knowledge for W, the augmented state vector $\boldsymbol{h}_{t-1}$ at any time $t-1$ is given by

$$\boldsymbol{h}_{t-1} = [\boldsymbol{x}\ w]^\intercal_{t-1}. \tag{8}$$

The prior predictive PDF of the hidden states $\boldsymbol{H}_{t|t-1}$ is given by

$$\boldsymbol{H}_{t|t-1} \sim \mathcal{N}(\boldsymbol{h}_t; \boldsymbol{\mu}^{\boldsymbol{H}}_{t|t-1}, \boldsymbol{\Sigma}^{\boldsymbol{H}}_{t|t-1}),$$

where, using Equations 1 & 7, the mean vector and the covariance matrix are given by

$$\begin{aligned}
\boldsymbol{\mu}^{\boldsymbol{H}}_{t|t-1} &= \begin{bmatrix} \mathbf{A}\boldsymbol{\mu}_{t-1|t-1} \\ 0 \end{bmatrix}_{t|t-1}, \\
\boldsymbol{\Sigma}^{\boldsymbol{H}}_{t|t-1} &= \begin{bmatrix} \mathbf{A}\boldsymbol{\Sigma}_{t-1|t-1}\mathbf{A}^\intercal + \mathbf{Q} & \boldsymbol{\Sigma}^{\boldsymbol{X}W} \\ (\boldsymbol{\Sigma}^{\boldsymbol{X}W})^\intercal & \mu^{\overline{W^2}} \end{bmatrix}_{t|t-1}.
\end{aligned} \tag{9}$$

The covariance term $\boldsymbol{\Sigma}^{\boldsymbol{X}W}_{t|t-1}$ between $\boldsymbol{X}_{t|t-1}$ and $W_{t|t-1}$ in Equation 9 is formulated as

$$\begin{aligned}
\boldsymbol{\Sigma}^{\boldsymbol{X}W}_{t|t-1} &= \operatorname{cov}(\mathbf{A}\boldsymbol{X}_{t-1|t-1} + \boldsymbol{W}_{t|t-1}, W_{t|t-1}), \\
&= \operatorname{cov}(\boldsymbol{W}, W)_{t|t-1},
\end{aligned} \tag{10}$$

where $\boldsymbol{W}_{t|t-1}$ is a vector of random variables representing the process error terms in the state vector $\boldsymbol{h}$. Moreover the hidden states $\boldsymbol{X}_{t-1|t-1}$ and the process error $W_{t|t-1}$ are assumed to be independent of each other. The mean and variance of $Y_{t|t-1} \sim \mathcal{N}(y_t; \mu_Y, \sigma^2_Y)$ are given by

$$\begin{aligned}
\mu_Y &= \mathbf{C}\boldsymbol{\mu}_{t|t-1} + \cancel{\mu_V}^{0}, \\
\sigma^2_Y &= \mathbf{C}\boldsymbol{\Sigma}_{t|t-1}\mathbf{C}^\intercal + \sigma^2_V,
\end{aligned}$$

given that $\boldsymbol{X}$ and V are assumed to be independent of each other. The covariance term $\boldsymbol{\Sigma}_{\boldsymbol{H}Y}$ between $\boldsymbol{H}_{t|t-1}$ and $Y_{t|t-1}$ is

$$\boldsymbol{\Sigma}_{\boldsymbol{H}Y} = \boldsymbol{\Sigma}^{\boldsymbol{H}}_{t|t-1}\mathbf{F}^\intercal_t,$$

where the observation matrix is $\mathbf{F}_t = [\mathbf{C}\ \ 0]$.

The inference for the parameter σ_W^2 requires two update steps; In the first step, the posterior PDF $f(\boldsymbol{h}_t|\boldsymbol{y}_{1:t})$ is estimated using the observation model defined in Equation 1 so that

$$f(\boldsymbol{h}_t|\boldsymbol{y}_{1:t}) = \frac{f(\boldsymbol{h}_t, y_t|\boldsymbol{y}_{1:t-1})}{f(y_t|\boldsymbol{y}_{1:t-1})} \approx \mathcal{N}(\boldsymbol{h}_t; \boldsymbol{\mu}_{t|t}^{\boldsymbol{H}}, \boldsymbol{\Sigma}_{t|t}^{\boldsymbol{H}}), \tag{11}$$

which we approximate by a Gaussian distribution with a posterior mean vector $\boldsymbol{\mu}_{t|t}^{\boldsymbol{H}}$ and a covariance matrix $\boldsymbol{\Sigma}_{t|t}^{\boldsymbol{H}}$ that are obtained using the predicted moments provided in Equation 9 and the Gaussian conditional equations [16, 33] such that

$$\begin{aligned} \boldsymbol{\mu}_{t|t}^{\boldsymbol{H}} &= \boldsymbol{\mu}_{t|t-1}^{\boldsymbol{H}} + \frac{\boldsymbol{\Sigma}_{\boldsymbol{H}Y}}{\sigma_Y^2}(y_t - \mu_Y), \\ \boldsymbol{\Sigma}_{t|t}^{\boldsymbol{H}} &= \boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{H}} - \frac{\boldsymbol{\Sigma}_{\boldsymbol{H}Y} \cdot \boldsymbol{\Sigma}_{\boldsymbol{H}Y}^{\intercal}}{\sigma_Y^2}. \end{aligned}$$

Now that we have the posterior PDF $f(w_t|\boldsymbol{y}_{1:t})$ from Equation 11, we move to the second update step where we use this new information of W at time t to update our current knowledge of $\overline{W^2}$. Figure 1 shows a graphical model representing the relationship between the random variables $Y_{t|t-1}$, $\boldsymbol{X}_{t|t-1}$, $W_{t|t-1}$, $W_{t|t-1}^2$ and $\overline{W_{t|t-1}^2}$. Note that while considering $\mu_W = 0$, the first moment of W^2 is equal to the second moment of W (under Lemma 1). In this case, the knowledge of W is fully defined by the knowledge of W^2, which is denoted in Figure 1 by an undirected solid line between the nodes W^2 and W. Following the structure depicted in Figure 1, the subsequent lemmas are provided for obtaining the posterior knowledge of $\overline{W_{t|t}^2}$.

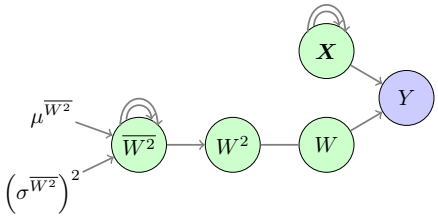


Figure 1: Illustration showing the graphical model for the online inference of the error variance parameter. The hidden and observed state variables are denoted by green and violet nodes. The double arrows on the nodes $\boldsymbol{X}$ and $\overline{W^2}$ represent that these variables are learnt recursively over time. For brevity, the subscript $t|t-1$ is dropped from each of the variables.

Lemma 3. *Considering the joint PDF of the random variables $Y_{t|t-1}$, $W_{t|t-1}^2$, and $\overline{W_{t|t-1}^2}$, and marginalizing out W^2 from the joint PDF, the posterior PDF of W^2 can be obtained by the following integral*

$$f(\overline{w^2}_t|\boldsymbol{y}_{1:t}) = \int f(w_t^2|\boldsymbol{y}_{1:t}) \cdot f(\overline{w^2}_t|w_t^2, \boldsymbol{y}_{1:t-1}) dw_t^2.$$

Proof. See Appendix B.3. □

Lemma 4. *The posterior mean and variance of* $W^2_{t|t}$ *are*

$$\begin{aligned}\mu^{W^2}_{t|t} &= (\mu^W_{t|t})^2 + (\sigma^W_{t|t})^2,\\ (\sigma^{W^2}_{t|t})^2 &= 2(\sigma^W_{t|t})^4 + 4(\sigma^W_{t|t})^2(\mu^W_{t|t})^2.\end{aligned}$$

Proof. See Appendix B.4. □

The Lemmas 3 & 4 are used for proving the following proposition.

Proposition 2. *The posterior mean and variance of* $\overline{W^2_{t|t}} \sim \mathcal{N}(\mu^{\overline{w^2}_t;\overline{W^2}}_{t|t}, (\sigma^{\overline{W^2}}_{t|t})^2)$ *are given by*

$$\begin{aligned}\mu^{\overline{W^2}}_{t|t} &= \mu^{\overline{W^2}}_{t|t-1} + k_t(\mu^{W^2}_{t|t} - \mu^{W^2}_{t|t-1}),\\ (\sigma^{\overline{W^2}}_{t|t})^2 &= (\sigma^{\overline{W^2}}_{t|t-1})^2 + k_t^2((\sigma^{W^2}_{t|t})^2 - (\sigma^{W^2}_{t|t-1})^2),\\ k &= \frac{(\sigma^{\overline{W^2}}_{t-1|t-1})^2}{(\sigma^{W^2}_{t|t-1})^2}.\end{aligned}$$

Proof. See Appendix B.5. □

Both the update steps 1 and 2 are employed recursively as observations are collected in order to first estimate the posterior knowledge of W and then use this to update our knowledge of the expected value of W^2, i.e., $\overline{W^2}$, which is a variable that is equal to σ^2_W, the parameter we seek to infer. All the steps performed in a particular time step t are summarized in Algorithm 1 as provided in Appendix D.

4 Multivariate Process Error

This section extends the mathematical formulation of the AGVI method for inferring error variance and covariance terms comprising a full $\mathbf{Q}$ matrix.

4.1 Problem Formulation

Let us consider $\mathtt{D}$ observed time series for which the global state vector is $\boldsymbol{x}_t = [\boldsymbol{x}^1_t\ \boldsymbol{x}^2_t\ \cdots\ \boldsymbol{x}^{\mathtt{D}}_t]^\intercal$, where $\boldsymbol{x}^j_t,\ \forall j \in \{1, 2, \cdots, \mathtt{D}\}$ refers to the concatenation of all $\mathtt{S}_j$ generic components for the j^{th} time series. Similarly, the vector of correlated process errors is assembled following $\boldsymbol{w}_t = [\boldsymbol{w}^1_t\ \boldsymbol{w}^2_t\ \cdots \boldsymbol{w}^{\mathtt{D}}_t]^\intercal$. The global transition, observation, process error covariance, and observation error covariance matrices are assembled block diagonally as

$$\begin{aligned}\mathbf{A} &= \text{blkdiag}[\mathbf{A}_1,\ \mathbf{A}_2,\ \cdots, \mathbf{A}_{\mathtt{D}}],\\ \mathbf{C} &= \text{blkdiag}[\mathbf{C}_1,\ \mathbf{C}_2,\ \cdots, \mathbf{C}_{\mathtt{D}}],\\ \mathbf{Q} &= \text{blkdiag}[\mathbf{Q}_1,\ \mathbf{Q}_2,\ \cdots, \mathbf{Q}_{\mathtt{D}}],\\ \mathbf{R} &= \text{blkdiag}[\mathbf{R}_1,\ \mathbf{R}_2,\ \cdots, \mathbf{R}_{\mathtt{D}}].\end{aligned}$$

Covariance matrices $\text{cov}(\boldsymbol{W}^k, \boldsymbol{W}^n)$ exists between the process errors $\boldsymbol{W}^k$ and $\boldsymbol{W}^n$ of the k^{th} and n^{th} time series respectively, where $k, n \in \{1, 2, \cdots, \mathtt{D}\}$. The process error covariance matrix $\mathbf{Q}$ can

be reformulated as follows

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_{1,2} & \cdots & \mathbf{Q}_{1,\mathtt{D}} \\ \vdots & \mathbf{Q}_2 & \cdots & \mathbf{Q}_{2,\mathtt{D}} \\ \vdots & \cdots & \ddots & \vdots \\ \text{sym.} & \cdots & \cdots & \mathbf{Q}_{\mathtt{D}} \end{bmatrix}, \tag{12}$$

where the covariance term $\text{cov}(\boldsymbol{W}^k, \boldsymbol{W}^n)$ is represented by $\mathbf{Q}_{k,n}$. The sub-matrices within the matrix $\mathbf{Q}_{k,n}$ in Equation 12 are themselves represented by $\text{cov}(\boldsymbol{W}^{jk}, \boldsymbol{W}^{mn}) = \mathbf{Q}_{jk,mn}$, where $j \in \{1, 2, \cdots, \mathtt{S}_j\}$ and $m \in \{1, 2, \cdots, \mathtt{S}_m\}$ are the j^{th} and m^{th} component of the k^{th} and n^{th} time series, respectively. As described in Section 3.1, each of the sub-matrices $\mathbf{Q}_{jk,mn}(\sigma^2_{jk}, \sigma^2_{mn}, \Delta t)$ can be represented as a function of the error variance parameters σ^2_{jk}, σ^2_{mn}, and Δt. Moreover, each of the elements within the sub-matrix $\mathbf{Q}_{jk,mn}$ is given by $\text{cov}(W^{ijk}, W^{lmn})$, which provides the covariance between the i^{th} process error term of the j^{th} component in the k^{th} time series, W^{ijk}, and the l^{th} process error term of the m^{th} component in the n^{th} time series, W^{lmn}. For example, let us consider two time series each modeled using a local trend (LT) component as described in Section 3.1. The global $\mathbf{Q}$ matrix is assembled block diagonally such that

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_{1,2} \\ \mathbf{Q}_{2,1} & \mathbf{Q}_2 \end{bmatrix},$$

where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are the process error covariance matrices associated with their individual local trend components and $\mathbf{Q}_{1,2}$ is the cross-covariance matrix between the process errors of the two time series. Each of these covariance matrices are defined as follows: $\mathbf{Q}_1 = \sigma^2_{\mathtt{LT}_1} \cdot \mathbf{J}$, $\mathbf{Q}_2 = \sigma^2_{\mathtt{LT}_2} \cdot \mathbf{J}$, and $\mathbf{Q}_{1,2} = \sigma_{\mathtt{LT}_{12}} \cdot \mathbf{J}$, where $\sigma^2_{\mathtt{LT}_1}$, and $\sigma^2_{\mathtt{LT}_2}$ are the two error variance terms for each of the LT component, $\sigma_{\mathtt{LT}_{12}}$ is the covariance term between the two process error random variables $W^{\mathtt{LT}_1}$ and $W^{\mathtt{LT}_2}$. For a constant acceleration kinematic model [3, 44], the matrix $\mathbf{J}$ is defined such that

$$\mathbf{J} = \begin{bmatrix} \frac{\Delta t^4}{4} & \frac{\Delta t^3}{2} \\ \frac{\Delta t^3}{2} & \Delta t^2 \end{bmatrix}.$$

Hence, for this case the terms to be inferred are: $\sigma^2_{\mathtt{LT}_1}$, $\sigma^2_{\mathtt{LT}_2}$, and $\sigma_{\mathtt{LT}_{12}}$. Similarly, for multiple time series, the goal is to infer one error variance term per time series along with the covariance terms for each pair of process error terms.

4.2 Methodology

Let us consider the multivariate process error term $\boldsymbol{w} = \begin{bmatrix} w^1 & w^2 & \cdots & w^i & \cdots & w^{\mathtt{D}} \end{bmatrix}^\intercal$, where w^i, $\forall i \in \{1,\ 2,\ \cdots, \mathtt{D}\}$ represents the process error term for the i^{th} time series for which the variance term has to be inferred. Given that the expected value of W is zero, the covariance term between the i^{th} and j^{th} process error is

$$\text{cov}(W^i, W^j) = \mathbb{E}[W^i W^j] - \cancelto{0}{\mathbb{E}[W^i]}\cancelto{0}{\mathbb{E}[W^j]} = \mathbb{E}[W^i W^j],$$

the covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ is given by

$$\boldsymbol{\Sigma}^{\boldsymbol{W}} = \begin{bmatrix} \mathbb{E}[(W^1)^2] & \mathbb{E}[W^1 W^2] & \cdots & \mathbb{E}[W^1 W^{\mathtt{D}}] \\ \vdots & \mathbb{E}[(W^2)^2] & \cdots & \mathbb{E}[W^2 W^{\mathtt{D}}] \\ \vdots & \cdots & \ddots & \vdots \\ \text{sym.} & \cdots & \cdots & \mathbb{E}[(W^{\mathtt{D}})^2] \end{bmatrix}, \tag{13}$$

where $\text{var}(W^i) = \mathbb{E}[(W^i)^2]$ is the error variance for the i^{th} time series, and $\text{cov}(W^i, W^j) = \mathbb{E}[W^i W^j]$ is the covariance term between the two process errors for the i^{th} and j^{th} time series. Similarly to the univariate process error, let us consider the approximation that each of the product terms $W^i W^j$ is a Gaussian random variable such that

$$W^i W^j \;\sim\; \mathcal{N}(w^i w^j; \mu^{W^i W^j}, (\sigma^{W^i W^j})^2), \tag{14}$$

where $\mathbb{E}[W^i W^j] = \mu^{W^i W^j}$ is the mean parameter and $\text{var}(W^i W^j) = (\sigma^{W^i W^j})^2$ is the variance. For $\mathtt{D}$ time series, there are a total of $\frac{\mathtt{D}\cdot(\mathtt{D}+1)}{2}$ product terms which are represented by the random vector $\boldsymbol{w^p} = [(w^1)^2\ (w^2)^2\ \cdots\ w^i w^j\ \cdots\ w^{\mathtt{D}} w^{\mathtt{D}-1}]^\intercal$ such that

$$\boldsymbol{W^p} \sim \mathcal{N}(\boldsymbol{w^p}; \boldsymbol{\mu^{W^p}}, \boldsymbol{\Sigma^{W^p}}), \tag{15}$$

where using Equation 14, the mean vector of $\boldsymbol{W^p}$ is given by

$$\boldsymbol{\mu^{W^p}} = \left[\mu^{(W^1)^2}\ \mu^{(W^2)^2}\ \cdots \mu^{(W^{\mathtt{D}})^2}\ \mu^{W^1 W^2}\ \cdots\ \mu^{W^{\mathtt{D}-1} W^{\mathtt{D}}}\right]^\intercal_{k\times 1}. \tag{16}$$

Similarly to Lemma 1, the covariance matrix $\boldsymbol{\Sigma^{W^p}}$ can be obtained in terms of the mean parameters in $\boldsymbol{\mu^{W^p}}$ such that

$$\boldsymbol{\Sigma^{W^p}} = \begin{bmatrix} 2(\mu^{(W^1)^2})^2 & 2(\mu^{W^1 W^2})^2 & \cdots & 2\mu^{W^1 W^{\mathtt{D}-1}}\mu^{W^1 W^{\mathtt{D}}} \\ \vdots & 2(\mu^{(W^2)^2})^2 & \cdots & 2\mu^{W^2 W^{\mathtt{D}-1}}\mu^{W^1 W^{\mathtt{D}}} \\ \vdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ \text{sym.} & \cdots & \cdots & \mu^{(W^{\mathtt{D}-1})^2}\mu^{(W^{\mathtt{D}})^2} + (\mu^{W^{\mathtt{D}-1} W^{\mathtt{D}}})^2 \end{bmatrix}_{k\times k},$$

where the variance $\text{var}(W^i W^j)$ and the covariance $\text{cov}(W^i W^j, W^k W^m)$ terms for the product of the errors $\forall i, j, k, m \in \{1,\ 2,\ \cdots, \mathtt{D}\}$ are obtained using the GMA equations presented in Appendix A. The mean vector $\boldsymbol{\mu^{W^p}}$ defined in Equation 16 is considered to be random and denoted by $\overline{\boldsymbol{W^p}}$ having a Gaussian PDF given by

$$\overline{\boldsymbol{W^p}} \sim \mathcal{N}(\overline{\boldsymbol{w^p}}; \boldsymbol{\mu^{\overline{W^p}}}, \boldsymbol{\Sigma^{\overline{W^p}}}), \tag{17}$$

where the vector $\overline{\boldsymbol{w^p}} = \left[\overline{(w^1)^2}\ \overline{(w^2)^2}\ \cdots\ \overline{(w^{\mathtt{D}})^2}\ \overline{w^1 w^2}\ \cdots\ \overline{w^{\mathtt{D}-1} w^{\mathtt{D}}}\right]^\intercal$. The mean vector and the covariance matrix of $\overline{\boldsymbol{W^p}}$ are

$$\boldsymbol{\mu^{\overline{W^p}}} = \left[\mu^{\overline{(W^1)^2}}\ \mu^{\overline{(W^2)^2}}\ \cdots\ \mu^{\overline{W^{\mathtt{D}-1} W^{\mathtt{D}}}}\right]^\intercal_{k\times 1},$$

$$\boldsymbol{\Sigma^{\overline{W^p}}} = \begin{bmatrix} (\sigma^{\overline{(W^1)^2}})^2 & 0 & \cdots & 0 \\ \vdots & (\sigma^{\overline{(W^2)^2}})^2 & \cdots & 0 \\ \vdots & \cdots & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ \text{sym.} & \cdots & \cdots & (\sigma^{\overline{W^{\mathtt{D}-1} W^{\mathtt{D}}}})^2 \end{bmatrix}_{k\times k},$$

where the random variables in $\overline{\boldsymbol{W^p}}$ are assumed to be independent from each other as shown by the covariance matrix $\boldsymbol{\Sigma^{\overline{W^p}}}$ where the off-diagonal terms are zero.

Using the hyperprior $\overline{\boldsymbol{W^p}}$ defined in Equation 17, the first objective is to obtain the covariance matrix $\boldsymbol{\Sigma^W}$ defined in Equation 13 by obtaining the prior predictive PDF of $\boldsymbol{W^p_{t|t-1}}$ as provided by the following lemma and proposition.

Lemma 5. *Using the transition model* $\boldsymbol{w}_t^{\boldsymbol{p}} = \boldsymbol{w}_{t-1}^{\boldsymbol{p}}$*, the prior predictive PDF of* $\boldsymbol{W}_{t|t-1}^{\boldsymbol{p}}$ *is given by*

$$\boldsymbol{W}_{t|t-1}^{\boldsymbol{p}} \sim \mathcal{N}(\boldsymbol{\mu}_{t|t-1}^{\boldsymbol{W^p}}, \boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{W^p}}),$$

where the mean terms in $\boldsymbol{\mu}_{t|t-1}^{\boldsymbol{W^p}}$*, and the variance and covariance terms in* $\boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{W^p}}$ *are given by*

$$\begin{aligned}
\mathbb{E}[W^i W^j] &= \mu^{\overline{W^i W^j}}, \\
\text{var}((W^i)^2) &= 3(\sigma^{\overline{(W^i)^2}})^2 + 2(\mu^{\overline{(W^i)^2}})^2, \\
\text{var}(W^i W^j) &= (\sigma^{\overline{W^i W^j}})^2 \\
&\quad + \frac{(\mu^{\overline{W^i W^j}})^2}{\mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\mu^{\overline{W^i W^j}})^2} \cdot (\sigma^{\overline{W^i W^j}})^2 \\
&\quad + \mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\mu^{\overline{W^i W^j}})^2, \\
\text{cov}(W^i W^j, W^l W^m) &= \mu^{\overline{W^i W^l}}\mu^{\overline{W^j W^m}} + \mu^{\overline{W^i W^m}}\mu^{\overline{W^j W^l}}.
\end{aligned}$$

Proof. See Appendix C.1. □

Proposition 3. *The prior predictive PDF of* $\boldsymbol{W}$ *has a zero mean vector and covariance matrix* $\boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{W}}$ *defined by*

$$\boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{W}} = \begin{bmatrix} \mu^{\overline{(W^1)^2}} & \mu^{\overline{W^1 W^2}} & \cdots & \mu^{\overline{W^1 W^{\mathtt{D}}}} \\ \vdots & \mu^{\overline{(W^2)^2}} & \cdots & \mu^{\overline{W^2 W^{\mathtt{D}}}} \\ \vdots & \cdots & \ddots & \vdots \\ sym. & \cdots & \cdots & \mu^{\overline{(W^{\mathtt{D}})^2}} \end{bmatrix}_{t|t-1}. \tag{18}$$

Proof. Using Lemma 5, the covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ for the prior predictive PDF of $\boldsymbol{W}$ is obtained by substituting the terms $\mathbb{E}[W^i W^j]$ in Equation 13 by the mean parameters of $\overline{\boldsymbol{W^p}}$, i.e., $\mu^{\overline{W^i W^j}}$. □

In order to maintain positive semi-definiteness of $\boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{W}}$ shown by Equation 18, the prior information is built from a random vector $\overrightarrow{\boldsymbol{L^W}}$ using Cholesky decomposition as shown by the following lemma.

Lemma 6. *Any* ij^{th} *element of* $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ *is obtained such that*

$$\mu^{\overline{W^i W^j}} = \mathbb{E}\left[\sum_{k=1}^{\mathtt{D}} L_{jk} L_{ki}\right],$$

where all elements of $\overrightarrow{\boldsymbol{L^W}}$ *are assumed to be Gaussian,* $L_{ij} \sim \mathcal{N}(\mu_{L_{ij}}, \sigma_{L_{ij}}^2)$*, and the expectation of the product terms are obtained using the GMA equations. Moreover, any covariance term between the random vectors* $\overrightarrow{\boldsymbol{L^W}}$ *and* $\overline{\boldsymbol{W^p}}$ *given by* $\boldsymbol{\Sigma}_{t|t-1}^{\overrightarrow{\boldsymbol{L^W}}\overline{\boldsymbol{W^p}}}$*, can be shown as*

$$\text{cov}(L_{ij}, \overline{W^i W^j}) = \text{cov}(L_{ij}, \sum_{k=1}^{\mathtt{D}} L_{jk} L_{ki}).$$

Proof. See Appendix C.2. □

Using the prior predictive PDF of $\boldsymbol{W}$, the next objective is to perform the prediction step. Let us consider the augmented vector of hidden states $\boldsymbol{h}_{t-1} = [\boldsymbol{x}_{t-1}^\intercal \ \boldsymbol{w}_{t-1}^\intercal]^\intercal$ such that the PDF of $\boldsymbol{H}_{t|t-1} \sim \mathcal{N}(\boldsymbol{h}_t, \boldsymbol{\mu}_{t|t-1}^{\boldsymbol{H}}, \boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{H}})$ has a mean vector $\boldsymbol{\mu}_{t|t-1}^{\boldsymbol{H}}$ and a covariance matrix $\boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{H}}$ defined by

$$\boldsymbol{\mu}_{t|t-1}^{\boldsymbol{H}} = \left[\boldsymbol{\mu}_{t|t-1}^\intercal \ \mathbf{0}\right]^\intercal, \tag{19}$$

$$\boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{H}} = \begin{bmatrix} \mathbf{A}\boldsymbol{\Sigma}_{t-1|t-1}\mathbf{A}^\intercal + \mathbf{Q} & \boldsymbol{\Sigma}^{\boldsymbol{XW}} \\ (\boldsymbol{\Sigma}^{\boldsymbol{XW}})^\intercal & \boldsymbol{\Sigma}^{\boldsymbol{W}} \end{bmatrix}_{t|t-1}, \tag{20}$$

where the covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ defined in Equation 18 is obtained using the prior knowledge of $\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}$ defined in the Cholesky space as stated in Lemma 6. Similarly to Equation 10, the covariance matrix between $\boldsymbol{X}$ and $\boldsymbol{W}$ is given by

$$\begin{aligned} \boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{XW}} &= \text{cov}(\boldsymbol{X}, \boldsymbol{W})_{t|t-1} \\ &= \text{cov}(\boldsymbol{X}, [W^1 \ W^2 \ \cdots \ W^{\mathtt{D}}]^\intercal)_{t|t-1}, \\ &= \text{cov}(\mathbf{A}\boldsymbol{X}_{t-1|t-1} + \boldsymbol{W}_{t|t-1}, [W^1 \ W^2 \ \cdots \ W^{\mathtt{D}}]_{t|t-1}^\intercal), \\ &= \text{cov}(\boldsymbol{W}, [W^1 \ W^2 \ \cdots \ W^{\mathtt{D}}]^\intercal)_{t|t-1}, \end{aligned}$$

where $\boldsymbol{W}_{t|t-1} = [W^1 \ W^2 \ \cdots \ W^{\mathtt{D}}]_{t|t-1}^\intercal$ is a vector of random variables that includes one process error term W from each of the $\mathtt{D}$ time series.

The inference for the covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ requires two update steps. Similarly to the univariate case shown in Section 3, the Gaussian conditional equations are used to perform the first update step to obtain the posterior PDF of $\boldsymbol{H}$ shown by

$$\boldsymbol{H}_{t|t} \sim \mathcal{N}(\boldsymbol{h}_t, \boldsymbol{\mu}_{t|t}^{\boldsymbol{H}}, \boldsymbol{\Sigma}_{t|t}^{\boldsymbol{H}}). \tag{21}$$

We now move to the second update step where we use the posterior PDF $f(\boldsymbol{w}_t|\boldsymbol{y}_{1:t})$ obtained from Equation 21, and the GMA equations to obtain the posterior PDF $f(\boldsymbol{w}_t^{\boldsymbol{p}}|\boldsymbol{y}_{1:t})$ such that

$$f(\boldsymbol{w}_t^{\boldsymbol{p}}|\boldsymbol{y}_{1:t}) = \mathcal{N}(\boldsymbol{w}_t^{\boldsymbol{p}}; \boldsymbol{\mu}_{t|t}^{\boldsymbol{W^p}}, \boldsymbol{\Sigma}_{t|t}^{\boldsymbol{W^p}}).$$

The posterior PDF of $\overline{\boldsymbol{W^p}}$ is defined using the following lemma.

Lemma 7. *The posterior mean, variance and covariance terms of* $\overline{\boldsymbol{W^p}}$ *are*

$$\begin{aligned} \boldsymbol{\mu}_{t|t}^{\overline{\boldsymbol{W^p}}} &= \boldsymbol{\mu}_{t|t-1}^{\overline{\boldsymbol{W^p}}} + \mathbf{K}_t(\boldsymbol{\mu}_{t|t}^{\boldsymbol{W^p}} - \boldsymbol{\mu}_{t|t-1}^{\boldsymbol{W^p}}), \\ \boldsymbol{\Sigma}_{t|t}^{\overline{\boldsymbol{W^p}}} &= \boldsymbol{\Sigma}_{t|t-1}^{\overline{\boldsymbol{W^p}}} + \mathbf{K}_t(\boldsymbol{\Sigma}_{t|t}^{\boldsymbol{W^p}} - \boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{W^p}})\mathbf{K}_t^\intercal, \\ \mathbf{K}_t &= \boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{W^p}\overline{\boldsymbol{W^p}}}(\boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{W^p}})^{-1}, \\ \boldsymbol{\Sigma}_{t|t-1}^{\boldsymbol{W^p}\overline{\boldsymbol{W^p}}} &= \boldsymbol{\Sigma}_{t|t-1}^{\overline{\boldsymbol{W^p}}}. \end{aligned}$$

Proof. See Appendix C.3. □

Using the updated knowledge of $\overline{\boldsymbol{W^p}}$ in Lemma 7, the posterior moments for $\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}$ in the Cholesky space is defined using the following proposition.

Proposition 4. *The posterior moments of* $\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}$ *are*

$$\begin{aligned} \boldsymbol{\mu}_{t|t}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}} &= \boldsymbol{\mu}_{t|t-1}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}} + \mathbf{K}_t^{\mathbf{L}}(\boldsymbol{\mu}_{t|t}^{\overline{\boldsymbol{W^p}}} - \boldsymbol{\mu}_{t|t-1}^{\overline{\boldsymbol{W^p}}}), \\ \boldsymbol{\Sigma}_{t|t}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}} &= \boldsymbol{\Sigma}_{t|t-1}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}} + \mathbf{K}_t^{\mathbf{L}}(\boldsymbol{\Sigma}_{t|t}^{\overline{\boldsymbol{W^p}}} - \boldsymbol{\Sigma}_{t|t-1}^{\overline{\boldsymbol{W^p}}})(\mathbf{K}_t^{\mathbf{L}})^\intercal, \\ \mathbf{K}_t^{\mathbf{L}} &= \boldsymbol{\Sigma}_{t|t-1}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}\overline{\boldsymbol{W^p}}}(\boldsymbol{\Sigma}_{t|t-1}^{\overline{\boldsymbol{W^p}}})^{-1}. \end{aligned}$$

Proof. The proposition 4 is derived using the Lemmas 5, 6, and 7. □

Both steps are employed recursively in order to estimate the elements of the covariance matrix $\mathbf{\Sigma}^{\boldsymbol{W}}$ and then use this to update our knowledge of the mean vector of $\boldsymbol{W^p}$, i.e., $\overline{\boldsymbol{W^p}}$. All the steps performed in a particular time step t are summarized in Algorithm 2 as provided in Appendix D.

5 Applied Examples

In this section, two case studies are presented to demonstrate the use of AGVI for online inference of error variance terms. Case study 1 focuses on the online estimation of error variance for a linear time-varying (LTV) model. The study includes statistical consistency tests to showcase the filter's optimality, t-tests to prove that the estimates are unbiased, empirical validation of uncertainty associated with error variance estimates, and analysis of the impact of $\frac{Q}{R}$ ratio on the posterior mean estimate $\mu_{\mathtt{T}|\mathtt{T}}$ of the error variance term. The AGVI method is also compared to two adaptive Kalman filtering (AKF) approaches: the *sliding window variational adaptive Kalman filter* (SWVAKF) [21] and the *measurement difference method* (MDM) [11, 13, 14, 26]. These AKF methods fall under different categories, with MDM being a correlation method and SWVAKF being a Bayesian method. Case study 2 presents a simulated multivariate random walk model with a full process error covariance matrix $\mathbf{Q}$ and a comparison of the AGVI method to the AKF methods.

5.1 Case Study 1 – Univariate Linear Time-Varying (LTV)

For this case study, a linear time-varying dynamic model is considered given by

$$\begin{aligned} x_t &= A_t x_{t-1} + w_t \quad w : W \sim \mathcal{N}(w; 0, \sigma_W^2), \\ y_t &= C_t x_t + v_t \qquad v : V \sim \mathcal{N}(v; 0, \sigma_V^2), \end{aligned} \tag{22}$$

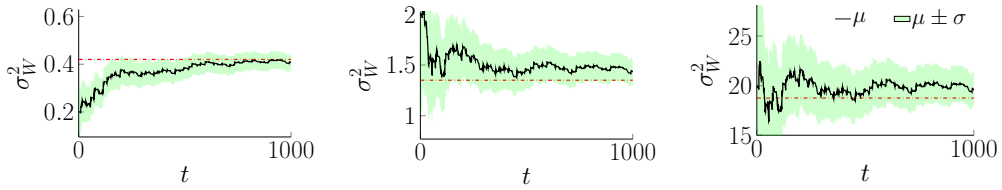
where the transition ($\mathbf{A}_t$) and observation ($\mathbf{C}_t$) equations at a time t are

$$\begin{aligned} A_t &= 0.8 - 0.1 \sin\left(\frac{7\pi t}{\mathtt{T}}\right), \\ C_t &= 1 - 0.99 \sin\left(\frac{100\pi t}{\mathtt{T}}\right), \end{aligned}$$

in which $\mathtt{T}$ represents the total number of time steps. The process error (w) and observation error (v) are Gaussian and are assumed to be independent at any time t. The process error variance σ_W^2 is unknown and needs to be inferred. The observation error variance σ_V^2 is assumed to be known and is equal to the true σ_W^2 which is randomly selected from the prior PDF of $\overline{W_{0|0}^2}$ such that

$$\overline{W_{0|0}^2} \sim \mathcal{N}(\overline{w_0^2}; \mu_{0|0}^{\overline{W^2}}, (\sigma_{0|0}^{\overline{W^2}})^2).$$

Three different true values for σ_W^2 are generated by considering different prior initialization for the pair of $\{\mu_{0|0}^{\overline{W^2}}, (\sigma_{0|0}^{\overline{W^2}})^2\}$ such that the three cases are (a) $\{\mu_{0|0}^{\overline{W^2}} = 0.2, (\sigma_{0|0}^{\overline{W^2}})^2 = 0.01\}$, (b) $\{\mu_{0|0}^{\overline{W^2}} = 2, (\sigma_{0|0}^{\overline{W^2}})^2 = 1\}$, and (c) $\{\mu_{0|0}^{\overline{W^2}} = 20, (\sigma_{0|0}^{\overline{W^2}})^2 = 100\}$.



(a) $\mu_{0|0}^{\overline{W^2}} = 0.2, (\sigma_{0|0}^{\overline{W^2}})^2 = 0.01$ (b) $\mu_{0|0}^{\overline{W^2}} = 2, (\sigma_{0|0}^{\overline{W^2}})^2 = 1$ (c) $\mu_{0|0}^{\overline{W^2}} = 20, (\sigma_{0|0}^{\overline{W^2}})^2 = 100$

Figure 2: Online estimation of the error variance term for each of the three cases for which the different prior initializations are (a) $\mu_{0|0}^{\overline{W^2}} = 0.2, (\sigma_{0|0}^{\overline{W^2}})^2 = 0.01$, (b) $\mu_{0|0}^{\overline{W^2}} = 2, (\sigma_{0|0}^{\overline{W^2}})^2 = 1$, and (c) $\mu_{0|0}^{\overline{W^2}} = 20, (\sigma_{0|0}^{\overline{W^2}})^2 = 100$. The true σ_W^2 value in each case is shown in red dashed line, while the estimated values and their $\pm 1\sigma$ uncertainty bound are shown in black and green shaded area.

Table 1: T-values computed in all three cases, i.e., (a) $\{\mu_{0|0}^{\overline{W^2}} = 0.2, (\sigma_{0|0}^{\overline{W^2}})^2 = 0.01\}$, (b) $\{\mu_{0|0}^{\overline{W^2}} = 2, (\sigma_{0|0}^{\overline{W^2}})^2 = 1\}$, and (c) $\{\mu_{0|0}^{\overline{W^2}} = 20, (\sigma_{0|0}^{\overline{W^2}})^2 = 100\}$.

t-values		
case (a)	case (b)	case (c)
−1.67	0.46	1.53
−1.29	−0.94	1.51
−0.87	1.26	−1.19
−0.79	1.21	1.03
−1.88	−0.64	−1.87

5.1.1 Online Estimation of Error Variance

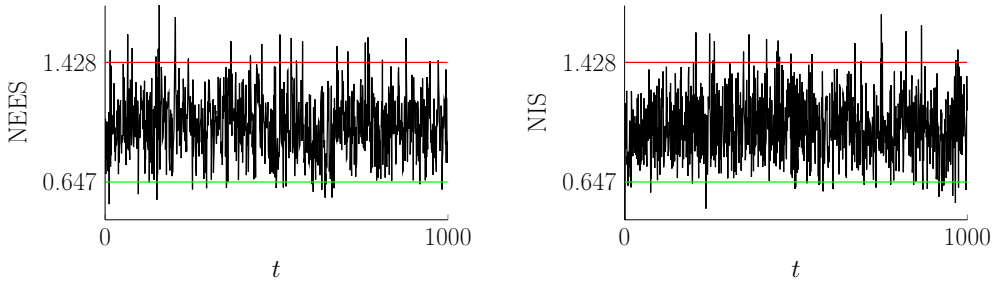
Data was simulated using the model specified in Equation 22 with a total of $\mathtt{T} = 1000$ observations. Figure 2 shows the online state estimation of the error variance term for each of the three cases. These results confirm that the method is able to perform online inference for different magnitudes of the error variance starting from arbitrary initial estimates.

Two-tailed t-tests were carried out to test the null hypothesis that the error variance estimates obtained using AGVI are equal to the true values. This empirical test, if accepted, proves that the estimates are unbiased. A 95% significance level is chosen for which the critical t-value is 1.96, given that the degrees of freedom are $\mathtt{T} - 1 = 999$. The t-test is carried out five times for each three cases. Table 1 shows the t-values where five different runs were carried out for each case. The results show that the computed t-values are within the bounds of the critical t-value, i.e., $-1.96 <$ t-value < 1.96, proving that the estimates are unbiased.

5.1.2 Statistical Consistency

The optimality of the filter is evaluated using two chi-square (χ^2) tests that rely on the *normalised estimation error squared* (NEES) and the *normalised innovation error squared* (NIS) values [3]. These tests are conducted using 50 random simulations. Considering a 95% confidence interval (C.I.)

and the degrees of freedom $v = \mathtt{X} = \mathtt{Y} = 1$, i.e., the size of the state and observation vector, the two-sided probability region is given by $\left[\chi^2_{50}(0.025)\ \chi^2_{50}(0.975)\right] = [32.3\ 71.4]$. By dividing the range by 50, we obtain the probability region for the average NEES and NIS values [0.647 1.428]. Figure 3 illustrates an example of the 95% region marked by the green and red lines for both the average NEES and NIS values in case (a). From the definition of the test, there should be approximately 5% of the total number of points outside the 95% region. The length of the time series for the case study is 1000 and hence, approximately 50 points should be outside the region. Table 2 presents the average number of points outside the probability region for both the NEES and NIS tests in all three cases where each of the 50 runs are carried out five times in order to compute the average value. The results verify that the filter is optimal and provide consistent estimates for the error variance term, given that the number of points outside the 95% probability region are in accordance to the theoretical results.



(a) NEES (b) NIS

Figure 3: Illustration showing the average normalized state estimation error squared (NEES) and the average normalized innovation squared (NIS) for the case study (a), i.e., $\{\mu^{\overline{W^2}}_{0|0} = 0.2, (\sigma^{\overline{W^2}}_{0|0})^2 = 0.01\}$, with its 95% probability region given by $[0.647, 1.428]$ is marked by the green and red lines.

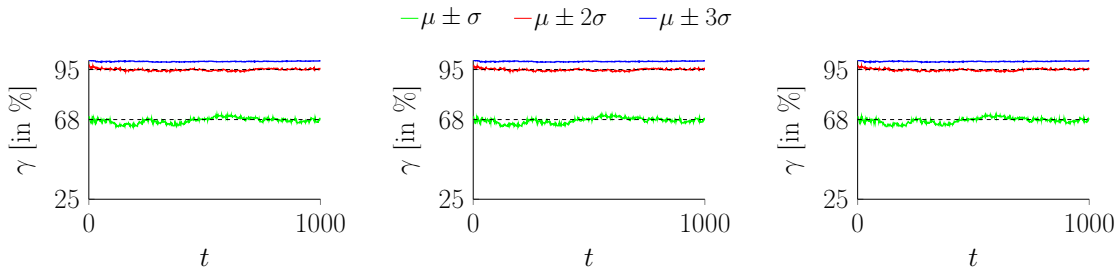
Table 2: Average number of points outside the 95% probability region for the NEES and NIS values in all the three cases, i.e., (a) $\{\mu^{\overline{W^2}}_{0|0} = 0.2, (\sigma^{\overline{W^2}}_{0|0})^2 = 0.01\}$, (b) $\{\mu^{\overline{W^2}}_{0|0} = 2, (\sigma^{\overline{W^2}}_{0|0})^2 = 1\}$, and (c) $\{\mu^{\overline{W^2}}_{0|0} = 20, (\sigma^{\overline{W^2}}_{0|0})^2 = 100\}$.

	Average number of points		
Criteria	case (a)	case (b)	case (c)
NEES	51	50.8	51.2
NIS	53	51.2	49.8

5.1.3 Statistical Consistency for the Variance of the Error Variance

In order to check the statistical consistency for the variance of the error variance term, we created 1000 simulated time series where the true values of the error variance in each time series is generated from the prior knowledge of $\overline{W^2_{0|0}}$. Figure 4 presents, for each time step, the percentage of realizations (γ) where the true value lies within the confidence interval (C.I.) for 1, 2, and 3 standard deviations

from the mean estimate in each of the three case studies. The results in Figure 4 show that the γ values match the theoretical C.I. quantities, i.e., $\{68, 95, 99\}\%$, for the Gaussian distribution supporting the hypothesis that the Gaussian PDF for the error variance is adequate at each time step.

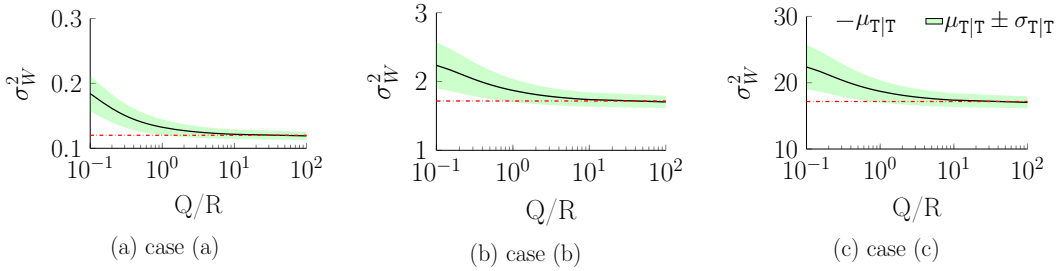


(a) case (a) (b) case (b) (c) case (c)

Figure 4: Empirical consistency check for the variance of the error variance estimate, where γ is the percentage of realizations where the true value lies within the three C.I. for the cases (a) $\{\mu_{0|0}^{\overline{W^2}} = 0.2, (\sigma_{0|0}^{\overline{W^2}})^2 = 0.01\}$, (b) $\{\mu_{0|0}^{\overline{W^2}} = 2, (\sigma_{0|0}^{\overline{W^2}})^2 = 1\}$, and (c) $\{\mu_{0|0}^{\overline{W^2}} = 20, (\sigma_{0|0}^{\overline{W^2}})^2 = 100\}$.

5.1.4 Impact of the $\frac{Q}{R}$ Ratio

Also, we noticed the effect of the $\frac{Q}{R} = \frac{\sigma_W^2}{\sigma_V^2}$ ratio on the estimation accuracy. Figure 5 shows the posterior mean estimate $\mu_{\mathtt{T}|\mathtt{T}}$ and the confidence interval $\mu_{\mathtt{T}|\mathtt{T}} \pm \sigma_{\mathtt{T}|\mathtt{T}}$ for the error variance after $\mathtt{T}$ time steps with respect to different $\frac{Q}{R}$ values for the three cases, where $\mathtt{T} = 1000$ is the total length of the time series. The results validate that the AGVI method is accurate for $\frac{Q}{R} \geq 10$. For $1 < \frac{Q}{R} < 10$, the estimates are accurate with negligible biases in comparison to the true values, whereas for $\frac{Q}{R} < 1$, the estimates are inaccurate with large biases. This phenomenon is explained by the fact that, given that the system is observable [7, 44], the Kalman gain has a higher value with an increase in the $\frac{Q}{R}$ ratio. As a result, the Kalman filter puts more weight on the measurements. Hence, we obtain a better mean and variance estimate of W by learning from each measurement which in turn provide better estimates for $\overline{W^2}$.



(a) case (a) (b) case (b) (c) case (c)

Figure 5: The posterior mean estimate and C.I of the error variance for different values of $\frac{Q}{R}$ for the cases (a) $\{\mu_{0|0}^{\overline{W^2}} = 0.2, (\sigma_{0|0}^{\overline{W^2}})^2 = 0.01\}$, (b) $\{\mu_{0|0}^{\overline{W^2}} = 2, (\sigma_{0|0}^{\overline{W^2}})^2 = 1\}$, and (c) $\{\mu_{0|0}^{\overline{W^2}} = 20, (\sigma_{0|0}^{\overline{W^2}})^2 = 100\}$. Note that the x-axis is in log-scale.

5.1.5 Comparison of AGVI with Adaptive Kalman Filters (AKF)

The AGVI method is compared with two existing adaptive Kalman filter (AKF) methods, namely SWVAKF [20] and MDM [11, 26]. Three cases are created for comparison, where the true values for the process error variances are (a) $\sigma_W^2 = 0.42$, (b) $\sigma_W^2 = 1.35$, and (c) $\sigma_W^2 = 18.75$, using the same dynamic model shown in Equation 22. In this case study, we use the publicly available codes for MDM [12] and SWVAKF [20] that only provides point estimates of $\mathbf{Q}$ and $\mathbf{R}$ matrices. As the codes are not available for obtaining the uncertainty associated with the error variance estimates, our study is restricted to analyzing the mean estimates, as made available by the authors. The MDM method involves no prior knowledge of the hidden states or the error variances, and the user-defined lag parameter is set to 1. For SWVAKF, the same values are used for filtering parameters as provided in the method's implementation code [21]. The prior knowledge for the hidden states in both AGVI and SWVAKF are set to $\mu_{0|0} = 0$ and $\sigma_{0|0}^2 = 100$.

Table 3 compares the average root mean square error (RMSE) values and computational time obtained using all three methods for the three cases, where the true values are (a) $\sigma_W^2 = 0.42$, (b) $\sigma_W^2 = 1.35$, and (c) $\sigma_W^2 = 18.75$. The results show that AGVI outperforms all methods in terms of predictive capacity. In comparison to SWVAKF, which is a Bayesian method, AGVI is more than two orders of magnitude faster, and compared to MDM, which is a correlation method, it is an order of magnitude faster. Thus, we conclude that AGVI provides unbiased and consistent estimates for the process error variance given that the observation error variance is known. The comparative study shows that AGVI provides better predictive capacity and computational speed than both AKF methods.

Table 3: Comparison of the average RMSE values and the computational time (s) obtained from each method in all three cases where the true values are (a) $\sigma_W^2 = 0.42$, (b) $\sigma_W^2 = 1.35$, and (c) $\sigma_W^2 = 18.75$. The results are averaged over five independent runs. Both the methods are picked from different AKF categories where AGVI and SWVAKF are Bayesian methods, MDM is a correlation method.

Category	Methods	RMSE			Time (s)
		$\sigma_W^2 = 0.42$	$\sigma_W^2 = 1.35$	$\sigma_W^2 = 18.75$	
Bayesian	AGVI	**0.043**	**0.083**	**2.06**	0.004
Bayesian	SWVAKF	0.067	0.138	3.18	2.638
Correlation	MDM	0.053	0.094	2.65	0.069

5.2 Case Study 2 – Multivariate Random Walk Model

This case study is conducted using five simulated datasets of 1000 time steps with a transition process error having a full covariance matrix $\mathbf{Q}$. The vector of hidden states $\boldsymbol{x}_t$ associated with the five time series is given by

$$\boldsymbol{x}_t = [x_t^{\mathtt{LL}_1}\ x_t^{\mathtt{LL}_2}\ x_t^{\mathtt{LL}_3}\ x_t^{\mathtt{LL}_4}\ x_t^{\mathtt{LL}_5}]^\intercal.$$

The state transition matrix $\mathbf{A}$ and the observation matrix $\mathbf{C}$ are defined as $\mathbf{A} = \boldsymbol{I}_5$, and $\mathbf{C} = \boldsymbol{I}_5$, The $\mathbf{Q}$ and $\mathbf{R}$ matrices are defined as

$$\mathbf{Q} = \begin{bmatrix} 1 & -0.3 & -0.2 & -0.1 & 0.25 \\ -0.3 & 3 & 0.35 & 0.4 & 0.45 \\ -0.2 & 0.35 & 4 & 0.5 & 0.55 \\ -0.1 & 0.4 & 0.5 & 0.8 & 0.6 \\ 0.25 & 0.45 & 0.55 & 0.6 & 2 \end{bmatrix},$$
$$\mathbf{R} = 0.1 \cdot \boldsymbol{I}_5,$$

where the off-diagonal covariance terms in the $\mathbf{Q}$ matrix are selected arbitrarily such that it is symmetric and semi positive-definite.

5.2.1 Online Estimation of Full Q Matrix

For AGVI, the prior knowledge for the augmented hidden states $\tilde{\boldsymbol{\mu}}_{0|0} = [\boldsymbol{\mu}_{0|0}; \boldsymbol{\mu}_{0|0}^{\overrightarrow{\boldsymbol{LW}}}]$ and $\tilde{\boldsymbol{\Sigma}}_{0|0} = \text{blkdiag}(\boldsymbol{\Sigma}_{0|0}, \boldsymbol{\Sigma}_{0|0}^{\overrightarrow{\boldsymbol{LW}}})$ are initialized by

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_{0|0} &= [\mathbf{0}_5^\intercal \;\; 2 \cdot \mathbf{1}_5^\intercal \;\; 0.8 \cdot \mathbf{1}_{10}^\intercal]^\intercal, \\ \tilde{\boldsymbol{\Sigma}}_{0|0} &= \text{diag}([\mathbf{1}_5^\intercal \;\; 0.5 \cdot \mathbf{1}_5^\intercal \;\; 0.5 \cdot \mathbf{1}_{10}^\intercal]), \end{aligned} \tag{23}$$

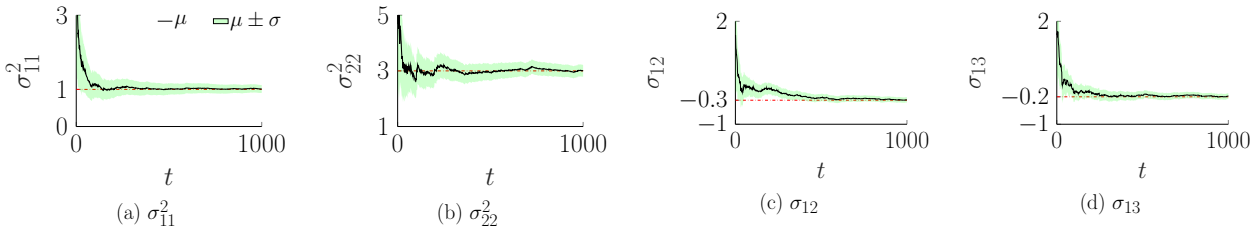
where $\mathbf{0}$ and $\mathbf{1}$ represent vector of zeros and ones, respectively. The mean vector and the covariance matrix for

$$\overrightarrow{\boldsymbol{LW}}_{0|0} = [L_{11} \; L_{22} \; \cdots L_{55} \; L_{12} \; \cdots L_{45}]_{0|0}^\intercal,$$

are given by

$$\begin{aligned} \boldsymbol{\mu}_{0|0}^{\overrightarrow{\boldsymbol{LW}}} &= [2 \cdot \mathbf{1}_5^\intercal \;\; 0.8 \cdot \mathbf{1}_{10}^\intercal]^\intercal, \\ \boldsymbol{\Sigma}_{0|0}^{\overrightarrow{\boldsymbol{LW}}} &= \text{diag}([\, 0.5 \cdot \mathbf{1}_5^\intercal \;\; 0.5 \cdot \mathbf{1}_{10}^\intercal]). \end{aligned} \tag{24}$$

Note that the matrix $\mathbf{R}$ is assumed to be known and only the matrix $\mathbf{Q}$ is inferred. Figure 6 shows the estimates obtained using AGVI for four elements of the $\mathbf{Q}$ matrix, namely σ_{11}^2, σ_{22}^2, σ_{12}, and σ_{13}, chosen arbitrarily. The plots for the remaining elements are provided in Appendix E.



(a) σ_{11}^2 (b) σ_{22}^2 (c) σ_{12} (d) σ_{13}

Figure 6: Online estimation of the error variance terms (a) σ_{11}^2 and (b) σ_{22}^2 and the covariance terms (c) σ_{12} and (d) σ_{13} from the full $\mathbf{Q}$ matrix compared to their true values marked by the dashed red line. The estimated values are shown by the black solid line and their $\pm 1\sigma$ uncertainty bound is shown using the green shaded region.

The two-tailed t-test was carried out to test the null hypothesis that the error variance estimates obtained using AGVI are equal to the true values and hence, unbiased. A 95% significance level was chosen for which the critical t-value is 1.96, given that the degree of freedom is $\mathtt{T} - 1 = 999$. Table 4 shows the average t-values for all the variance terms. The results show that the computed t-values are within the bounds of the critical t-value, i.e., $-1.96 <$ t-value < 1.96, proving that the estimates are unbiased.

Table 4: Average t-values for all the variance terms in the full $\mathbf{Q}$ matrix. Five independent runs were carried out. The variance terms and the covariance terms are represented by σ_{ii}^2 and $\sigma_{ij},\ \forall i, j \in 1, \cdots, \mathtt{D}$, respectively.

	Variance terms														
	σ_{11}^2	σ_{12}	σ_{13}	σ_{14}	σ_{15}	σ_{22}^2	σ_{23}	σ_{24}	σ_{25}	σ_{33}^2	σ_{34}	σ_{35}	σ_{44}^2	σ_{45}	σ_{55}^2
t-values	0.6002	−0.0736	0.1006	0.1380	−0.2828	0.3387	0.2157	0.2527	−0.1177	−0.0423	−0.0981	−0.2424	0.0692	−0.0855	−0.0287

5.2.2 Statistical Consistency

The normalized innovation square (NIS) metric is used to check the consistency of the estimator. The two-sided probability region for a 95% C.I. having five degrees of freedom, i.e., the size of the observation vector $\mathtt{Y}$, is [0.831 12.833]. Considering that the total length of the training set is 1000, the theoretical 5% value for the number of acceptable points outside the 95% C.I. is 50. The different prior initialization are chosen such that $\boldsymbol{\mu}_{0|0}^{\overrightarrow{\boldsymbol{LW}}} = [\alpha \cdot \mathbf{1}_5^\intercal \ \beta \cdot \mathbf{1}_{10}^\intercal]^\intercal$, where $\alpha = \{1.5 : 0.1 : 2\}$ and $\beta = \{0.5 : 0.1 : 1\}$ while considering the same covariance matrix as defined in Equation 24. Table 5 presents the average number of points outside the probability region for the different prior initialization of $\overrightarrow{\boldsymbol{LW}}_{0|0}$ where the average value is computed using five simulated datasets for each combination of $\{\alpha, \beta\}$. The results show that, on average, there are ≈ 56 points that lie outside the 95% probability region, which is comparable to the theoretical value of 50. This verifies that the filter is optimal and provides consistent estimates for the error variance and covariance terms of the full $\mathbf{Q}$ matrix.

Table 5: Average number of points outside the 95% probability region for the different prior initialization of $\overrightarrow{\boldsymbol{LW}}_{0|0}$. Each column presents the average value computed using the five simulated datasets for one combination of $\{\alpha, \beta\}$.

	$\{1.5, 0.5\}$	$\{1.6, 0.6\}$	$\{1.7, 0.7\}$	$\{1.8, 0.8\}$	$\{1.9, 0.9\}$	$\{2, 1\}$	Mean
NIS	54.8	54.6	53.8	53.8	53.6	53.6	54.03

5.2.3 Comparison with Adaptive Kalman Filters (AKF)

For this case study, the AGVI method is compared to SWVAKF and MDM. For the SWVAKF method, the hidden states are initialized similarly to Equation 23, where the mean vector is $\boldsymbol{\mu}_{0|0} = \mathbf{0}_5$ and the covariance matrix is $\boldsymbol{\Sigma}_{0|0} = \boldsymbol{I}_5$. The same values are used for filtering parameters as provided

in the method's implementation code [21]. The MDM method involves no prior knowledge of the hidden states or error variances, and the user-defined lag parameter is set to 1. Table 6 shows the average RMSE values for estimating some of the elements chosen arbitrarily from the $\mathbf{Q}$ matrix as well as the average computational time for each method. The results show that AGVI outperforms other methods in terms of predictive capacity for most of the variance and covariance terms. In comparison to SWVAKF which is a Bayesian method, it is more than an order of magnitude faster. The MDM is an order of magnitude faster than AGVI but the current implementation code only provides point estimates, whereas AGVI estimates the mean as well as the variance of the error variance terms. The results for both case studies were obtained using the AGVI MATLAB library [8].

Table 6: Comparison of the average RMSE values and the computational time for each method. Each of the methods are picked from different AKF categories where AGVI and SWVAKF are Bayesian methods, MDM is a correlation method. The variance terms and the covariance terms are represented by σ_{ii}^2 and $\sigma_{ij},\ \forall i,j \in 1,\cdots,\mathtt{D}$, respectively.

Methods	Variance terms															Time (s)
	σ_{11}^2	σ_{12}	σ_{13}	σ_{14}	σ_{15}	σ_{22}^2	σ_{23}	σ_{24}	σ_{25}	σ_{33}^2	σ_{34}	σ_{35}	σ_{44}^2	σ_{45}	σ_{55}^2	
AGVI	0.0766	0.0409	**0.0160**	0.0476	**0.0139**	**0.0468**	0.2097	**0.0004**	**0.0578**	0.1653	0.0528	**0.0123**	0.0251	**0.0115**	**0.0422**	1.4
SWVAKF	0.0246	0.0789	0.1364	0.0857	0.1023	0.5607	**0.0754**	0.0995	0.2539	1.1391	0.1661	0.1366	**0.0130**	0.0265	0.2304	9.4
MDM	**0.0034**	**0.0197**	0.0675	**0.0437**	0.0629	0.4431	0.5396	0.0581	0.1306	**0.0347**	**0.0009**	0.0736	0.0413	0.0609	0.1412	**0.03**

6 Conclusion

The approximate Gaussian variance inference (AGVI) method proposed in this paper is an analytically tractable Bayesian inference method that provides many advantages. Firstly, it allows for online inference of the process error variance and covariance terms in the full $\mathbf{Q}$ matrix as Gaussian random variables. Secondly, it provides accurate, unbiased, and statistically consistent estimates of the mean and uncertainties associated with the error variance terms at each time step. Thirdly, it employs a closed-form square-root filtering technique using the Cholesky decomposition to maintain the estimated $\mathbf{Q}$ matrix as positive semi-definite.

The case study 1 shows the application of the AGVI method for a linear time-varying model where the univariate process error variance was inferred starting from different prior initializations. The t-test proves that the estimates are unbiased. The statistical consistency tests verify that the filter is optimal and that the AGVI method provide consistent estimates for the mean as well as the uncertainties associated with the error variance term. In comparison to the existing adaptive Kalman filter (AKF) methods, the AGVI method provides a better predictive capacity in all cases. In comparison to SWVAKF, which is a Bayesian method, AGVI is more than two orders of magnitude faster, and compared to MDM, which is a correlation method, it is an order of magnitude faster. The case study 2 shows the application of AGVI for a multivariate random walk model with a full $\mathbf{Q}$ matrix and compares its performance with the two AKF methods. The results show that AGVI outperforms all methods in terms of predictive capacity for most of the variance and covariance terms, and yields statistically consistent estimates. Hence, the proposed method is capable of online estimation of error variances in regard to state-space models involving multiple time series.

Acknowledgment

This project was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC), Hydro-Québec (HQ), Hydro-Québec's Research Institute (IREQ), Institute For Data Valorization (IVADO). The authors thank Luong Ha Nguyen, Post doc, Department of Civil, Geologic and Mining Engineering, Polytechnique Montréal for his help in the project.

References

[1] Tohid Ardeshiri, Emre Özkan, Umut Orguner, and Fredrik Gustafsson. Approximate Bayesian smoothing with unknown process and measurement noise covariances. *IEEE Signal Processing Letters*, 22(12):2450–2454, 2015.

[2] Akbar Assa and Konstantinos N Plataniotis. Adaptive Kalman filtering by covariance sampling. *IEEE Signal Processing Letters*, 24(9):1288–1292, 2017.

[3] Yaakov Bar-Shalom, X Rong Li, and Thiagalingam Kirubarajan. *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.

[4] Vinay A Bavdekar, Anjali P Deshpande, and Sachin C Patwardhan. Identification of process and measurement noise covariance for state and parameter estimation using extended Kalman filter. *Journal of Process Control*, 21(4):585–601, 2011.

[5] Pierre R Belanger. Estimation of noise covariance matrices for a linear time-varying stochastic process. *Automatica*, 10(3):267–275, 1974.

[6] Alessandro Brumana and Luigi Piroddi. A multi-tone central divided difference frequency tracker with adaptive process noise covariance tuning. *International Journal of Adaptive Control and Signal Processing*, 34(7):877–900, 2020.

[7] Arthur E Bryson. Applied optimal control: Optimization. *Estimization and Control*, 2, 1975.

[8] Bhargob Deka and James-A Goulet. AGVI. `https://github.com/CivML-PolyMtl/AGVI`, 2023. Accessed: 2023-05-13.

[9] Bhargob Deka, Luong Ha Nguyen, Saeid Amiri, and James-A Goulet. The Gaussian multiplicative approximation for state-space models. *Structural Control and Health Monitoring*, page e2904, 2021.

[10] Jindřich Duník, Oliver Kost, and Ondřej Straka. Design of measurement difference autocovariance method for estimation of process and measurement noise covariances. *Automatica*, 90:16–24, 2018.

[11] Jindřich Duník, Oliver Kost, Ondřej Straka, and Erik Blasch. Covariance estimation and gaussianity assessment for state and measurement noise. *Journal of Guidance, Control, and Dynamics*, 43(1):132–139, 2020.

[12] Jindřich Duník, Ondřej Straka, and Oliver" Kost. The Identification and Decision Making Research Group (IDM). `https://idm.kky.zcu.cz/sw.html`. Accessed: 2023-05-09.

[13] Jindřich Duník, Ondřej Straka, and Oliver Kost. Measurement difference autocovariance method for noise covariance matrices estimation. 2016 IEEE 55th Conference on Decision and Control (CDC), pages 365–370. IEEE, 2016.

[14] Jindřich Duník, Ondřej Straka, Oliver Kost, and Jindřich Havlík. Noise covariance matrices in state-space models: A survey and comparison of estimation methods,Äîpart I. *International Journal of Adaptive Control and Signal Processing*, 31(11):1505–1543, 2017.

[15] Paul Fearnhead. Markov chain Monte Carlo, sufficient statistics, and particle filters. *Journal of Computational and Graphical Statistics*, 11(4):848–862, 2002.

[16] A Gelman, B Carlin, S Stern, B Dunson, A Vehtari, and B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[17] James-A Goulet. *Probabilistic machine learning for civil engineers*. MIT Press, 2020.

[18] James-A Goulet, Luong Ha Nguyen, and Saeid Amiri. Tractable approximate Gaussian inference for Bayesian neural networks. *Journal of Machine Learning Research*, 22(251):1–23, 2021.

[19] Yulong Huang, Yonggang Zhang, Zhemin Wu, Ning Li, and Jonathon Chambers. A novel adaptive Kalman filter with inaccurate process and measurement noise covariance matrices. *IEEE Transactions on Automatic Control*, 63(2):594–601, 2017.

[20] Yulong Huang, Fengchi Zhu, Guangle Jia, and Yonggang Zhang. Implementation codes for the paper a slide window variational adaptive Kalman filter. `https://www.researchgate.net/publication/342466130_Implementation_codes_for_the_paper_A_Slide_Window_Variational_Adaptive_Kalman_Filter`, 06 2020.

[21] Yulong Huang, Fengchi Zhu, Guangle Jia, and Yonggang Zhang. A slide window variational adaptive Kalman filter. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67(12):3552–3556, 2020.

[22] Nikolas Kantas, Arnaud Doucet, Sumeetpal S Singh, Jan Maciejowski, and Nicolas Chopin. On particle methods for parameter estimation in state-space models. *Statistical Science*, 30(3):328–351, 2015.

[23] R Kashyap. Maximum likelihood identification of stochastic linear systems. *IEEE Transactions on Automatic Control*, 15(1):25–34, 1970.

[24] Steven M Kay. *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.

[25] Thaleia Kontoroupi and Andrew W Smyth. Online noise identification for joint state and parameter estimation of nonlinear systems. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 2(3):B4015006, 2016.

[26] Oliver Kost, Jindřich Duník, and Ondřej Straka. Measurement difference method: A universal tool for noise identification. *IEEE Transactions on Automatic Control*, 68(3):1792–1799, 2022.

[27] X Rong Li and Yaakov Bar-Shalom. A recursive multiple model approach to noise identification. *IEEE Transactions on Aerospace and Electronic Systems*, 30(3):671–684, 1994.

[28] Jirong Ma, Hua Lan, Zengfu Wang, Xuezhi Wang, Quan Pan, and Bill Moran. Improved adaptive Kalman filter with unknown process noise covariance. 2018 21st International Conference on Information Fusion, pages 1–5. IEEE, 2018.

[29] Peter Matisko and Vladimír Havlena. Noise covariance estimation for Kalman filter tuning using Bayesian approach and Monte Carlo. *International Journal of Adaptive Control and Signal Processing*, 27(11):957–973, 2013.

[30] R Mehra. On the identification of variances and adaptive Kalman filtering. *IEEE Transactions on Automatic Control*, 15(2):175–184, 1970.

[31] R Mehra. On-line identification of linear dynamic systems with applications to Kalman filtering. *IEEE Transactions on Automatic Control*, 16(1):12–21, 1971.

[32] Raman Mehra. Approaches to adaptive filtering. *IEEE Transactions on Automatic Control*, 17(5):693–698, 1972.

[33] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

[34] Vincent Mussot, Guillaume Mercère, Thibault Dairay, Vincent Arvis, and Jérémy Vayssettes. Noise covariance matrix estimation with subspace model identification for Kalman filtering. *International Journal of Adaptive Control and Signal Processing*, 35(4):591–611, 2021.

[35] Kenneth Myers and BD Tapley. Adaptive sequential estimation with unknown noise statistics. *IEEE Transactions on Automatic Control*, 21(4):520–523, 1976.

[36] H Nguyen and J-A Goulet. Real-time anomaly detection with Bayesian dynamic linear models. *Structural Control and Health Monitoring*, page e2404, 2019.

[37] Brian J Odelson, Murali R Rajamani, and James B Rawlings. A new autocovariance least-squares method for estimating noise covariances. *Automatica*, 42(2):303–308, 2006.

[38] Lenka Pavelková and Miroslav Kárnỳ. State and parameter estimation of state-space model with entry-wise correlated uniform noise. *International Journal of Adaptive Control and Signal Processing*, 28(11):1189–1205, 2014.

[39] P Sage and W Husa. Algorithms for sequential adaptive estimation of prior statistics. 1969 IEEE Symposium on Adaptive Processes (8th) Decision and Control, pages 61–61. IEEE, 1969.

[40] Simo Särkkä. *Bayesian filtering and smoothing*. Cambridge University Press, 2013.

[41] Simo Särkkä and Jouni Hartikainen. Non-linear noise adaptive Kalman filtering via variational Bayes. 2013 IEEE International Workshop on Machine Learning for Signal Processing, pages 1–6. IEEE, 2013.

[42] Simo Sarkka and Aapo Nummenmaa. Recursive noise adaptive Kalman filtering by variational Bayesian approximations. *IEEE Transactions on Automatic control*, 54(3):596–600, 2009.

[43] Robert H Shumway, David S Stoffer, and David S Stoffer. *Time series analysis and its applications*. Springer, 2000.

[44] Dan Simon. *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006.

[45] Geir Storvik. Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions on Signal Processing*, 50(2):281–289, 2002.

[46] J Wang. Stochastic Modeling for Real-Time Kinematic GPS/GLONASS Positioning. *Navigation*, 46(4):297–305, 1999.

[47] J Zhou and RH Luecke. Estimation of the covariances of the process noise and measurement noise for a linear discrete dynamic system. *Computers & Chemical Engineering*, 19(2):187–195, 1995.

A Gaussian Multiplication Approximation (GMA)

Consider $\boldsymbol{X} = [X_1\ X_2\ X_3\ X_4]^\intercal$, a vector of Gaussian random variables such that $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Using the Gaussian moment generating function or 2^{nd} order Taylor series expansion, the following equations hold for the product of any two Gaussian random variables such that

$$\begin{aligned}
\mathbb{E}[X_1X_2] &= \mu_1\mu_2 + \text{cov}(X_1, X_2), && (25)\\
\text{var}(X_1X_2) &= \sigma_1^2\sigma_2^2 + \text{cov}(X_1, X_2)^2 \\
&\quad + 2\text{cov}(X_1, X_2)\mu_1\mu_2 && (26)\\
&\quad + \sigma_1^2\mu_2^2 + \sigma_2^2\mu_1^2, \\
\text{cov}(X_3, X_1X_2) &= \text{cov}(X_1, X_3)\mu_2 + \text{cov}(X_2, X_3)\mu_1, && (27)\\
\text{cov}(X_1X_2, X_3X_4) &= \text{cov}(X_1, X_3)\text{cov}(X_2, X_4) && (28)\\
&\quad + \text{cov}(X_1, X_4)\text{cov}(X_2, X_3) \\
&\quad + \text{cov}(X_1, X_3)\mu_2\mu_4 \\
&\quad + \text{cov}(X_1, X_4)\mu_2\mu_3 + \text{cov}(X_2, X_3)\mu_1\mu_4 \\
&\quad + \text{cov}(X_2, X_4)\mu_1\mu_3.
\end{aligned}$$

These equations can be obtained either using moment generating functions [18] or second-order Taylor series expansion [9].

B Proofs for Univariate Process Error

B.1 Proof for Lemma 1

Proof. Given that W is Gaussian and has a zero mean, the moments of W can be derived using a Gaussian moment generating function so that

$$\begin{aligned}
\mu_W &= \mathbb{E}[W] = 0, \\
\sigma_W^2 &= \mathbb{E}[(W - \mu_W)^2] = \mathbb{E}[W^2] - \mathbb{E}[W]^2, && (29)\\
&= \mathbb{E}[W^2], \\
\mathbb{E}[W^4] &= 3\mathbb{E}[W^2]^2, && (30)
\end{aligned}$$

where using Equations 29 & 30, we can define the variance of W^2 such that

$$\text{var}(W^2) = \mathbb{E}[(W^4)] - \mathbb{E}[W^2]^2 = 2\mathbb{E}[W^2]^2. \tag{31}$$

If we make the approximation that $W^2 \sim \mathcal{N}(w^2; \mu^{W^2}, (\sigma^{W^2})^2)$ is a Gaussian random variable, then the PDF can be fully defined by its mean and variance,

$$\begin{aligned}
\mu^{W^2} &= \mathbb{E}[W^2], \\
(\sigma^{W^2})^2 &= \text{var}(W^2) = 2\mathbb{E}[W^2]^2,
\end{aligned}$$

where by using Equation 31, the variance $\text{var}(W^2)$ can also be expressed in terms of the expected value $\mathbb{E}[W^2]$. Hence, the PDF of W^2 only depends on the unknown hyper parameter μ^{W^2} such that

$$\begin{aligned}
f(w^2|\mu^{W^2}, (\sigma^{W^2})^2) &\equiv f(w^2|\mu^{W^2}), \\
&= \mathcal{N}(w^2, \mu^{W^2}, 2(\mu^{W^2})^2). && (32)
\end{aligned}$$

□

B.2 Proof for Lemma 2

Proof. Let us consider that the mean parameter μ^{W^2} is described by the random variable $\overline{W^2}$: $\overline{w^2} \in (0, \infty)$ for which

$$f(\overline{w^2}) \sim \mathcal{N}(\overline{w^2}; \mu^{\overline{W^2}}, (\sigma^{\overline{W^2}})^2). \tag{33}$$

Using (32) and (33), we can rewrite the PDF of W^2 as

$$f(w^2|\overline{w^2}) = \mathcal{N}(w^2; \overline{w^2}, 2(\overline{w^2})^2). \tag{34}$$

Using the acyclic graph in Figure 1, the joint PDF of $Y_{t|t-1}$, $\boldsymbol{X}_{t|t-1}$, $W^2_{t|t-1}$, and $\overline{W^2_{t|t-1}}$is shown by

$$\begin{aligned} f(y_t, x_t, w_t^2, \overline{w^2}_t|\boldsymbol{y}_{1:t-1}) = f(y_t|x_t, w_t^2) \cdot f(x_t|\boldsymbol{y}_{1:t-1}) \\ \cdot f(w_t^2|\overline{w^2}_t) \cdot f(\overline{w^2}_t|\boldsymbol{y}_{1:t-1}). \end{aligned} \tag{35}$$

Using Equation 34 and marginalizing $Y_{t|t-1}$, $\boldsymbol{X}_{t|t-1}$, and $\overline{W^2_{t|t-1}}$ from the joint PDF defined in Equation 35, the prior predictive PDF of $W^2_{t|t-1}$ is

$$\begin{aligned} f(w_t^2|\boldsymbol{y}_{1:t-1}) &= \int f(w_t^2|\overline{w^2}_t) \cdot f(\overline{w^2}_t|\boldsymbol{y}_{1:t-1}) d\overline{w^2}_t, \\ &= \int \mathcal{N}(w_t^2; \overline{w^2}_t, 2(\overline{w^2}_t)^2) \cdot \mathcal{N}(\overline{w^2}_t; \mu^{\overline{W^2}}_{t-1|t-1}, (\sigma^{\overline{W^2}}_{t-1|t-1})^2) d\overline{w^2}_t. \end{aligned} \tag{36}$$

The integration in Equation 36 can be solved in closed-form. The equivalent formulation to obtain this is to represent the Gaussian random variables W^2 and $\overline{W^2}$ in terms of the standard Gaussian variable ϵ and ζ shown by

$$W^2 = \overline{W^2} + \sqrt{2}\,\overline{W^2}\epsilon, \quad \epsilon \sim \mathcal{N}(0,1) \tag{37}$$

$$\overline{W^2} = \mu^{\overline{W^2}} + \sigma^{\overline{W^2}}\zeta, \quad \zeta \sim \mathcal{N}(0,1). \tag{38}$$

Using Equations 37 & (38), the mean and variance of the prior predictive PDF of $W^2_{t|t-1}$ are given by

$$\begin{aligned} \mathbb{E}[W^2_{t|t-1}] &= \mathbb{E}[\overline{W^2_{t|t-1}}] + \sqrt{2}\cancelto{0}{\mathbb{E}[\overline{W^2_{t|t-1}}\epsilon]}, \\ &= \mu^{\overline{W^2}}_{t-1|t-1}, \end{aligned} \tag{39}$$

$$\begin{aligned} \mathrm{var}(W^2_{t|t-1}) &= \mathrm{var}(\overline{W^2_{t|t-1}}) + 2\,\mathrm{var}(\overline{W^2_{t|t-1}}\epsilon), \\ &= (\sigma^{\overline{W^2}}_{t-1|t-1})^2 + 2(\mathrm{var}(\overline{W^2_{t|t-1}}) \cdot \cancelto{1}{\mathrm{var}(\epsilon)} \\ &+ \cancelto{1}{\mathrm{var}(\epsilon)} \cdot \mathbb{E}[\overline{W^2_{t|t-1}}]^2), \\ &= 3(\sigma^{\overline{W^2}}_{t-1|t-1})^2 + 2(\mu^{\overline{W^2}}_{t-1|t-1})^2, \end{aligned} \tag{40}$$

where the term $\mathrm{var}(\overline{W^2_{t|t-1}}\epsilon)$ in Equation 40 is obtained using the GMA equations,

$$\mathrm{var}(\overline{W^2_{t|t-1}}\epsilon) = \mathrm{var}(\overline{W^2_{t|t-1}}) \cdot \mathrm{var}(\epsilon) + \mathrm{var}(\epsilon) \cdot \mathbb{E}[\overline{W^2_{t|t-1}}]^2.$$

Using Equations 29 & 39, the error variance term is given by

$$\sigma_W^2 = \mu^{\overline{W^2}}_{t-1|t-1}. \tag{41}$$

□

B.3 Proof for Lemma 3

Proof. Let us consider the joint PDF of $Y_{t|t-1}$, $W^2_{t|t-1}$, and $\overline{W^2}_{t|t-1}$as shown by Figure 1,

$$\begin{aligned} f(y_t, w_t^2, \overline{w^2}_t|\boldsymbol{y}_{1:t-1}) &= f(y_t|w_t^2) \cdot f(w_t^2|\overline{w^2}_t) \cdot f(\overline{w^2}_t|\boldsymbol{y}_{1:t-1}). \\ &= f(y_t|w_t^2) \cdot \frac{f(w_t^2|\overline{w^2}_t) \cdot f(\overline{w^2}_t|\boldsymbol{y}_{1:t-1})}{f(w_t^2|\boldsymbol{y}_{1:t-1})} \cdot f(w_t^2|\boldsymbol{y}_{1:t-1}), \\ &= f(y_t|w_t^2) \cdot \frac{f(w_t^2, \overline{w^2}_t|\boldsymbol{y}_{1:t-1})}{f(w_t^2|\boldsymbol{y}_{1:t-1})} \cdot f(w_t^2|\boldsymbol{y}_{1:t-1}), \\ &= f(y_t|w_t^2) \cdot f(\overline{w^2}_t|w_t^2, \boldsymbol{y}_{1:t-1}) \cdot f(w_t^2|\boldsymbol{y}_{1:t-1}). \end{aligned} \tag{42}$$

By dividing both sides in Equation 42 by $f(y_t|\boldsymbol{y}_{1:t-1})$ we obtain

$$\begin{aligned} \frac{f(y_t, w_t^2, \overline{w^2}_t|\boldsymbol{y}_{1:t-1})}{f(y_t|\boldsymbol{y}_{1:t-1})} &= \frac{f(y_t|w_t^2) \cdot f(w_t^2|\boldsymbol{y}_{1:t-1})}{f(y_t|\boldsymbol{y}_{1:t-1})} \cdot f(\overline{w^2}_t|w_t^2, \boldsymbol{y}_{1:t-1}), \\ f(w_t^2, \overline{w^2}_t|\boldsymbol{y}_{1:t}) &= f(w_t^2|\boldsymbol{y}_{1:t}) \cdot f(\overline{w^2}_t|w_t^2, \boldsymbol{y}_{1:t-1}), \end{aligned} \tag{43}$$

By marginalizing out W^2 from the joint PDF defined in Equation 43, the posterior PDF of $\overline{W^2}_{t|t}$ is obtained such that

$$f(\overline{w^2}_t|\boldsymbol{y}_{1:t}) = \int f(w_t^2|\boldsymbol{y}_{1:t}) \cdot f(\overline{w^2}_t|w_t^2, \boldsymbol{y}_{1:t-1}) dw_t^2. \tag{44}$$

□

B.4 Proof for Lemma 4

Proof. Using the GMA equations in Section A, the posterior PDF of W^2 can be shown by

$$f(w_t^2|\boldsymbol{y}_{1:t}) \sim \mathcal{N}(w_t^2; \mu^{W^2}_{t|t}, (\sigma^{W^2}_{t|t})^2), \tag{45}$$

where the mean and variance of $W^2_{t|t}$ are

$$\begin{aligned} \mu^{W^2}_{t|t} &= (\mu^W_{t|t})^2 + (\sigma^W_{t|t})^2, \\ (\sigma^{W^2}_{t|t})^2 &= 2(\sigma^W_{t|t})^4 + 4(\sigma^W_{t|t})^2(\mu^W_{t|t})^2. \end{aligned}$$

□

B.5 Proof for Proposition 2

Proof. Given that the prior predictive PDF of both W^2 and $\overline{W^2}$ are Gaussian, the joint multivariate Gaussian PDF $f(\overline{w^2}_t, w_t^2|\boldsymbol{y}_{1:t-1})$ is shown by

$$f(\overline{w^2}_t, w_t^2|\boldsymbol{y}_{1:t-1}) = \mathcal{N}\Big(\begin{pmatrix} \overline{w^2}_t \\ w_t^2 \end{pmatrix}; \boldsymbol{\mu}^{\overline{W^2}, W^2}_{t|t-1}, \boldsymbol{\Sigma}^{\overline{W^2}, W^2}_{t|t-1}\Big), \tag{46}$$

having a mean vector $\boldsymbol{\mu}^{\overline{W^2}, W^2}_{t|t-1}$ and a covariance matrix $\boldsymbol{\Sigma}^{\overline{W^2}, W^2}_{t|t-1}$ defined as

$$\begin{aligned} \boldsymbol{\mu}^{\overline{W^2}, W^2}_{t|t-1} &= \begin{bmatrix} \mu^{\overline{W^2}}_{t|t-1} & \mu^{W^2}_{t|t-1} \end{bmatrix}^\intercal, \\ \boldsymbol{\Sigma}^{\overline{W^2}, W^2}_{t|t-1} &= \begin{bmatrix} (\sigma^{\overline{W^2}})^2_{t|t-1} & \text{cov}(\overline{W^2}, W^2)_{t|t-1} \\ \text{cov}(W^2, \overline{W^2})_{t|t-1} & (\sigma^{W^2})^2_{t|t-1} \end{bmatrix}, \end{aligned} \tag{47}$$

and where using the transition model $w_t^2 = w_{t-1}^2$, the mean and the variance of $W_{t|t-1}^2 = W_{t-1|t-1}^2$ are given by Equations 39 & 40. The covariance term $\text{cov}(W^2, \overline{W^2})_{t|t-1}$ between $W_{t|t-1}^2$ and $\overline{W_{t|t-1}^2}$ in Equation 47 is obtained using Equations 33 & 37, and the GMA equations from Section A so that

$$\begin{aligned}
\text{cov}(W_{t|t-1}^2, \overline{W_{t|t-1}^2}) &= \text{cov}(W^2, \overline{W^2})_{t|t-1}, \\
&= \text{cov}(\overline{W^2} + \sqrt{2}\,\overline{W^2}\epsilon, \overline{W^2})_{t|t-1}, \\
&= \text{var}(\overline{W^2})_{t|t-1} + \sqrt{2}\text{cov}(\overline{W^2}\epsilon, \overline{W^2})_{t|t-1}, \\
&= \text{var}(\overline{W^2})_{t|t-1} + \sqrt{2}(\text{cov}(\overline{W^2}, \overline{W^2})\cancelto{0}{\mathbb{E}[\epsilon]} \\
&+ \cancelto{0}{\text{cov}(\epsilon, \overline{W^2})}\mathbb{E}[\overline{W^2}]), \\
&= (\sigma_{t-1|t-1}^{\overline{W^2}})^2.
\end{aligned}$$

Given that the joint Gaussian PDF is defined as shown by Equation 46, the Gaussian conditional properties are used to obtain the conditional PDF $f(\overline{w^2}_t|w_t^2, \boldsymbol{y}_{1:t-1})$ which is part of the integrand shown in Equation 44,

$$f(\overline{w^2}_t|w_t^2, \boldsymbol{y}_{1:t-1}) = \mathcal{N}(\overline{w^2}_t; \mu_{t|t-1}^{\overline{W^2}|W^2}, (\sigma_{t|t-1}^{\overline{W^2}|W^2})^2), \tag{48}$$

for which the conditional mean and variance are

$$\mu_{t|t-1}^{\overline{W^2}|W^2} = \mu_{t|t-1}^{\overline{W^2}} + k_t(w_t^2 - \mu_{t|t-1}^{W^2}), \tag{49}$$

$$(\sigma_{t|t-1}^{\overline{W^2}|W^2})^2 = (\sigma_{t|t-1}^{\overline{W^2}})^2 - k_t^2(\sigma_{t|t-1}^{W^2})^2, \tag{50}$$

$$\begin{aligned}
k_t &= \frac{\text{cov}(W_{t|t-1}^2, \overline{W_{t|t-1}^2})}{(\sigma_{t|t-1}^{W^2})^2}, \\
&= \frac{(\sigma_{t-1|t-1}^{\overline{W^2}})^2}{(\sigma_{t|t-1}^{W^2})^2}.
\end{aligned} \tag{51}$$

Using Equations 48 & 45, Equation 44 is rewritten as

$$f(\overline{w^2}_t|\boldsymbol{y}_{1:t}) = \int \mathcal{N}(\overline{w^2}_t; \mu_{t|t-1}^{\overline{W^2}|W^2}, (\sigma_{t|t-1}^{\overline{W^2}|W^2})^2) \cdot \mathcal{N}(w_t^2; \mu_{t|t}^{W^2}, (\sigma_{t|t}^{W^2})^2) dw_t^2. \tag{52}$$

Equation 52 can be solved in closed-form having a Gaussian PDF with a random mean, i.e., $\mu_{t|t-1}^{\overline{W^2}|W^2}$, and a constant variance, i.e., $(\sigma_{t|t-1}^{\overline{W^2}|W^2})^2$, shown by Equations 49 & 50. Hence, the PDF $f(\overline{w^2}_t|\boldsymbol{y}_{1:t})$ is also Gaussian such that

$$f(\overline{w^2}_t|\boldsymbol{y}_{1:t}) = \mathcal{N}(\overline{w^2}_t; \mu_{t|t}^{\overline{W^2}}, (\sigma_{t|t}^{\overline{W^2}})^2),$$

for which the posterior mean and the variance can be computed using the Kalman gain defined in Equation 51 as

$$\begin{aligned}
\mu_{t|t}^{\overline{W^2}} &= \mathbb{E}\left[\mu_{t|t-1}^{\overline{W^2}} + k_t(W_{t|t}^2 - \mu_{t|t-1}^{W^2})\right], \\
&= \mu_{t|t-1}^{\overline{W^2}} + k_t(\mu_{t|t}^{W^2} - \mu_{t|t-1}^{W^2}), \\
(\sigma_{t|t}^{\overline{W^2}})^2 &= (\sigma_{t|t-1}^{\overline{W^2}})^2 - k_t^2(\sigma_{t|t-1}^{W^2})^2 + k_t^2\text{var}(W_{t|t}^2), \\
&= (\sigma_{t|t-1}^{\overline{W^2}})^2 + k_t^2((\sigma_{t|t}^{W^2})^2 - (\sigma_{t|t-1}^{W^2})^2).
\end{aligned} \tag{53}$$

□

C Proofs for Multivariate Process Error

C.1 Proof for Lemma 5

As described by Lemma B.2 for the univariate process error, the expected value $\mathbb{E}[(W^i)^2]$ and the variance terms $\text{var}((W^i)^2), \forall i \in \{1, 2, \cdots, \mathtt{D}\}$ for the prior predictive PDF of $\boldsymbol{W^p}$ are given by

$$\begin{aligned}\mathbb{E}[(W^i)^2] &= (\mu^{\overline{(W^i)^2}}),\\ \text{var}((W^i)^2) &= 3(\sigma^{\overline{(W^i)^2}})^2 + 2(\mu^{\overline{(W^i)^2}})^2.\end{aligned}$$

Using the GMA equations, the mean and variance term of W^iW^j are

$$\mathbb{E}[W^iW^j] = \text{cov}(W^i, W^j) = \overline{w^iw^j}, \tag{54}$$

$$\begin{aligned}\text{var}(W^iW^j) &= \text{var}(W^i)\text{var}(W^j) + \text{cov}(W^i, W^j)^2,\\ &= \mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\overline{w^iw^j})^2,\end{aligned} \tag{55}$$

where using Equation 41 from Proof B.2, $\text{var}(W^i) = \mu^{\overline{(W^i)^2}}$. Using Equations 54 and 55, the Gaussian random variable $W^iW^j \sim \mathcal{N}(w^iw^j; \overline{w^iw^j}, \mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\overline{w^iw^j})^2)$ can be represented in terms of its standard Gaussian variable ϵ by

$$w^iw^j = \overline{w^iw^j} + \sqrt{\mu^{\overline{(W^i)^2}} \cdot \mu^{\overline{(W^j)^2}} + (\overline{w^iw^j})^2} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1).$$

The moments of the prior predictive PDF of W^iW^j are given by

$$\begin{aligned}\mathbb{E}[W^iW^j] &= \mathbb{E}[\overline{W^iW^j}] + \mathbb{E}\left[\sqrt{\mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\overline{W^iW^j})^2} \cdot \epsilon\right], \quad (\to 0)\\ &= \mu^{\overline{W^iW^j}},\end{aligned} \tag{56}$$

$$\begin{aligned}\text{var}(W^iW^j) &= \text{var}(\overline{W^iW^j}) + \text{var}(\sqrt{\mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\overline{W^iW^j})^2} \cdot \epsilon),\\ &= \text{var}(\overline{W^iW^j}) + \text{var}(\sqrt{\mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\overline{W^iW^j})^2}),\\ &\quad + \mathbb{E}\left[\sqrt{\mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\overline{W^iW^j})^2}\right]^2,\end{aligned} \tag{57}$$

where using GMA equations the term $\text{var}(\sqrt{\mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\overline{W^iW^j})^2} \cdot \epsilon)$ is obtained by

$$\text{var}(\sqrt{\mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\overline{W^iW^j})^2} \cdot \epsilon) = \text{var}(\sqrt{\mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\overline{W^iW^j})^2}) + \mathbb{E}\left[\sqrt{\mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\overline{W^iW^j})^2}\right]^2.$$

In order to simplify the notation in Equation 57, let us consider $u = \overline{w^iw^j}$ and $t(u) = \sqrt{\mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\overline{w^iw^j})^2}$, so that using 1^{st} order Taylor series expansion we get

$$\mathbb{E}[t(u)]^2 = t(\mathbb{E}[U])^2 = \mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\mu^{\overline{W^iW^j}})^2, \tag{58}$$

$$\begin{aligned}\text{var}(t(u)) &= (t'(\mathbb{E}[U]))^2 \cdot \text{var}(U),\\ &= \frac{(\mu^{\overline{W^iW^j}})^2}{\mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\mu^{\overline{W^iW^j}})^2} \cdot (\sigma^{\overline{W^iW^j}})^2.\end{aligned} \tag{59}$$

Hence, combining Equations 57, 58, and 59 we get

$$\begin{aligned}\text{var}(W^iW^j) &= (\sigma^{\overline{W^iW^j}})^2 + \frac{(\mu^{\overline{W^iW^j}})^2}{\mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\mu^{\overline{W^iW^j}})^2} \cdot (\sigma^{\overline{W^iW^j}})^2 \\ &\quad + \mu^{\overline{(W^i)^2}}\mu^{\overline{(W^j)^2}} + (\mu^{\overline{W^iW^j}})^2.\end{aligned}$$

Using the GMA equations, and Equation 56, the covariance between the product terms W^iW^j and W^lW^m, $i, j, l, m \; \forall \in \{1, 2, \cdots \mathtt{D}\}$, is given by

$$\begin{aligned}\text{cov}(W^iW^j, W^lW^m) &= \text{cov}(W^i, W^l)\text{cov}(W^j, W^m) + \text{cov}(W^i, W^m)\text{cov}(W^j, W^l), \\ &= \mathbb{E}[W^iW^l]\mathbb{E}[W^jW^m] + \mathbb{E}[W^iW^m]\mathbb{E}[W^jW^l], \\ &= \mu^{\overline{W^iW^l}}\mu^{\overline{W^jW^m}} + \mu^{\overline{W^iW^m}}\mu^{\overline{W^jW^l}}.\end{aligned}$$

C.2 Proof for Lemma 6

Proof. The covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ in Equation 13 can be reformulated in terms of the random variables in $\overline{\boldsymbol{W}^{\boldsymbol{p}}}$ given by

$$\boldsymbol{\Sigma}^{\boldsymbol{W}} = \begin{bmatrix} \overline{(W^1)^2} & \overline{W^1W^2} & \cdots & \overline{W^1W^{\mathtt{D}}} \\ \vdots & \overline{(W^2)^2} & \cdots & \overline{W^2W^{\mathtt{D}}} \\ \vdots & \cdots & \ddots & \vdots \\ \text{sym.} & \cdots & \cdots & \overline{(W^{\mathtt{D}})^2} \end{bmatrix}_{t|t-1}, \tag{60}$$

where using Equation 54, $\mathbb{E}[W^iW^j] = \overline{W^iW^j}, \forall i, j \in \{1,\ 2,\ \cdots, \mathtt{D}\}$. Let us consider $\boldsymbol{L}^{\boldsymbol{W}}$ is an upper triangular random matrix such that

$$\boldsymbol{L}^{\boldsymbol{W}} = \begin{bmatrix} L_{11} & L_{12} & \cdots & L_{1\mathtt{D}} \\ 0 & L_{22} & \cdots & L_{2\mathtt{D}} \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & 0 & L_{\mathtt{DD}} \end{bmatrix}, \tag{61}$$

where each of the term is assumed to be a Gaussian random variable given by $L_{ij} \sim \mathcal{N}(\mu_{L_{ij}}, \sigma^2_{L_{ij}})$. The elements of $\boldsymbol{L}^{\boldsymbol{W}}$ can be arranged in a random vector,

$$\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}} = [L_{11}\ L_{22}\ L_{\mathtt{DD}}\ L_{12}\ \cdots L_{ij}\ \cdots L_{\mathtt{D\text{-}1\,D}}]^\intercal,$$

such that $\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}$ is a Gaussian random vector given by

$$\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}} \sim \mathcal{N}(\boldsymbol{\mu}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}}, \boldsymbol{\Sigma}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}}), \tag{62}$$

where $\boldsymbol{\mu}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}}$ and $\boldsymbol{\Sigma}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}}$ are the mean vector and the covariance matrix of $\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}$. Let us reproduce $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ using Equation 61 such that

$$\boldsymbol{\Sigma}^{\boldsymbol{W}} = (\boldsymbol{L}^{\boldsymbol{W}})^\intercal \boldsymbol{L}^{\boldsymbol{W}},$$

where each element $\overline{W^iW^j}$ of $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ defined in Equation 60 is obtained using matrix multiplication so that

$$\overline{W^iW^j} = \sum_{k=1}^{\mathtt{D}} L_{jk}L_{ki}, \forall i, j \in \{1, \cdots, \mathtt{D}\},$$

where using Equation 62 and the GMA equations we can determine the expected value, the variance, and the covariance terms of any element $\overline{W^i W^j}$ as follows,

$$\mathbb{E}[\overline{W^i W^j}] = \mathbb{E}\left[\sum_{k=1}^{\mathtt{D}} L_{jk} L_{ki}\right], \quad \text{var}(\overline{W^i W^j}) = \text{var}(\sum_{k=1}^{\mathtt{D}} L_{jk} L_{ki}). \tag{63}$$

Using Equation 63, the elements of the prior predictive PDF of $\boldsymbol{W}$ defined in Proposition 3 can be computed as

$$\mu^{\overline{W^i W^j}} = \mathbb{E}\left[\sum_{k=1}^{\mathtt{D}} L_{jk} L_{ki}\right].$$

Similarly, the covariance between the random matrices, $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ and $\boldsymbol{L}^{\boldsymbol{W}}$, is equivalent to finding the covariance between the random vectors $\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}$ and $\overline{\boldsymbol{W}^{\boldsymbol{p}}}$ given by $\boldsymbol{\Sigma}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}\overline{\boldsymbol{W}^{\boldsymbol{p}}}}_{t|t-1}$, where any covariance term is obtained by

$$\text{cov}(L_{ij}, \overline{W^i W^j}) = \text{cov}(L_{ij}, \sum_{k=1}^{\mathtt{D}} L_{jk} L_{ki}).$$

□

C.3 Proof for Lemma 7

Proof. The prior knowledge of $\overline{\boldsymbol{W}^{\boldsymbol{p}}}$ is updated by employing the prior predictive $\boldsymbol{W}^{\boldsymbol{p}}_{t|t-1}$ and the posterior PDF $\boldsymbol{W}^{\boldsymbol{p}}_{t|t}$ such that

$$f(\overline{\boldsymbol{w}^{\boldsymbol{p}}_t}|\boldsymbol{y}_{1:t}) \;\; = \;\; \mathcal{N}(\overline{\boldsymbol{w}^{\boldsymbol{p}}_t}; \boldsymbol{\mu}^{\overline{\boldsymbol{W}^{\boldsymbol{p}}}}_{t|t}, \boldsymbol{\Sigma}^{\overline{\boldsymbol{W}^{\boldsymbol{p}}}}_{t|t}),$$

where using Equations 51-53, the posterior mean, variance and covariance terms of $\overline{\boldsymbol{W}^{\boldsymbol{p}}}$ are

$$\begin{aligned}
\boldsymbol{\mu}^{\overline{\boldsymbol{W}^{\boldsymbol{p}}}}_{t|t} &= \boldsymbol{\mu}^{\overline{\boldsymbol{W}^{\boldsymbol{p}}}}_{t|t-1} + \mathbf{K}_t(\boldsymbol{\mu}^{\boldsymbol{W}^{\boldsymbol{p}}}_{t|t} - \boldsymbol{\mu}^{\boldsymbol{W}^{\boldsymbol{p}}}_{t|t-1}),\\
\boldsymbol{\Sigma}^{\overline{\boldsymbol{W}^{\boldsymbol{p}}}}_{t|t} &= \boldsymbol{\Sigma}^{\overline{\boldsymbol{W}^{\boldsymbol{p}}}}_{t|t-1} + \mathbf{K}_t(\boldsymbol{\Sigma}^{\boldsymbol{W}^{\boldsymbol{p}}}_{t|t} - \boldsymbol{\Sigma}^{\boldsymbol{W}^{\boldsymbol{p}}}_{t|t-1})\mathbf{K}_t^\intercal,\\
\mathbf{K}_t &= \boldsymbol{\Sigma}^{\boldsymbol{W}^{\boldsymbol{p}}\overline{\boldsymbol{W}^{\boldsymbol{p}}}}_{t|t-1}(\boldsymbol{\Sigma}^{\boldsymbol{W}^{\boldsymbol{p}}}_{t|t-1})^{-1},\\
\boldsymbol{\Sigma}^{\boldsymbol{W}^{\boldsymbol{p}}\overline{\boldsymbol{W}^{\boldsymbol{p}}}}_{t|t-1} &= \boldsymbol{\Sigma}^{\overline{\boldsymbol{W}^{\boldsymbol{p}}}}_{t|t-1}.
\end{aligned}$$

□

D Algorithms for Univariate and Multivariate Process Errors

Algorithm 1 One-time step of the proposed AGVI method for univariate process error

Input: $\boldsymbol{\mu}_{t-1|t-1}$, $\boldsymbol{\Sigma}_{t-1|t-1}$, $\mu^{\overline{W^2}}_{t-1|t-1}$, $(\sigma^{\overline{W^2}}_{t-1|t-1})^2$, y_t, $\mathbf{A}$, $\mathbf{C}$, $\mathbf{Q}$, and σ^2_V

Prior knowledge for the error variance parameter:

$\sigma^2_W = \mu^{\overline{W^2}}_{t-1|t-1}$

Prediction Step:

$$\boldsymbol{\mu}^{\boldsymbol{H}}_{t|t-1} = \begin{bmatrix} \mathbf{A}\boldsymbol{\mu}_{t-1|t-1} \\ 0 \end{bmatrix}_{t|t-1}, \quad \boldsymbol{\Sigma}^{\boldsymbol{H}}_{t|t-1} = \begin{bmatrix} \mathbf{A}\boldsymbol{\Sigma}_{t-1|t-1}\mathbf{A}^\intercal + \mathbf{Q} & \boldsymbol{\Sigma}^{\boldsymbol{X}W} \\ (\boldsymbol{\Sigma}^{\boldsymbol{X}W})^\intercal & \mu^{\overline{W^2}} \end{bmatrix}_{t|t-1},$$

$$\mu_Y = \mathbf{C}\boldsymbol{\mu}_{t|t-1}, \quad \sigma^2_Y = \mathbf{C}\boldsymbol{\Sigma}_{t|t-1}\mathbf{C}^\intercal + \sigma^2_V, \quad \boldsymbol{\Sigma}_{\boldsymbol{H}Y} = \boldsymbol{\Sigma}^{\boldsymbol{H}}_{t|t-1}\mathbf{F}^\intercal_t$$

1^{st} **Update Step:**

$$\boldsymbol{\mu}^{\boldsymbol{H}}_{t|t} = \boldsymbol{\mu}^{\boldsymbol{H}}_{t|t-1} + \frac{\boldsymbol{\Sigma}_{\boldsymbol{H}Y}}{\sigma^2_Y}(y_t - \mu_Y), \quad \boldsymbol{\Sigma}^{\boldsymbol{H}}_{t|t} = \boldsymbol{\Sigma}^{\boldsymbol{H}}_{t|t-1} - \frac{\boldsymbol{\Sigma}_{\boldsymbol{H}Y}\cdot\boldsymbol{\Sigma}^\intercal_{\boldsymbol{H}Y}}{\sigma^2_Y}$$

Posterior Moments for W^2:

$$\mu^{W^2}_{t|t} = (\mu^W_{t|t})^2 + (\sigma^W_{t|t})^2,$$

$$(\sigma^{W^2}_{t|t})^2 = 2(\sigma^W_{t|t})^4 + 4(\sigma^W_{t|t})^2(\mu^W_{t|t})^2$$

2^{nd} **Update Step:**

$$\mu^{\overline{W^2}}_{t|t} = \mu^{\overline{W^2}}_{t|t-1} + k_t(\mu^{W^2}_{t|t} - \mu^{W^2}_{t|t-1}), \quad (\sigma^{\overline{W^2}}_{t|t})^2 = (\sigma^{\overline{W^2}}_{t|t-1})^2 + k^2_t((\sigma^{W^2}_{t|t})^2 - (\sigma^{W^2}_{t|t-1})^2),$$

$$k_t = \frac{(\sigma^{\overline{W^2}}_{t-1|t-1})^2}{(\sigma^{W^2}_{t|t-1})^2}$$

return $\boldsymbol{\mu}_{t|t}$, $\boldsymbol{\Sigma}_{t|t}$, $\mu^{\overline{W^2}}_{t|t}$, and $(\sigma^{\overline{W^2}}_{t|t})^2$

Algorithm 2 One-time step of the AGVI method for multivariate process errors

Input: $\boldsymbol{\mu}_{t-1|t-1}$, $\boldsymbol{\Sigma}_{t-1|t-1}$, $\boldsymbol{\mu}^{\overrightarrow{\boldsymbol{LW}}}_{t-1|t-1}$, $\boldsymbol{\Sigma}^{\overrightarrow{\boldsymbol{LW}}}_{t-1|t-1}$, $\boldsymbol{y}_t$, $\mathbf{A}$, $\mathbf{C}$, $\mathbf{Q}$, and $\mathbf{R}$

Prior Predictive PDF of $\boldsymbol{W}_{t|t-1} \sim \mathcal{N}(\boldsymbol{w}_t; \mathbf{0}_{t|t-1}, \boldsymbol{\Sigma}^{\boldsymbol{W}}_{t|t-1})$:

Any ij^{th} element of $\boldsymbol{\Sigma}^{\boldsymbol{W}}_{t|t-1}$ is obtained using $\mu^{\overline{W^i W^j}} = \mathbb{E}\left[\sum_{k=1}^{\mathtt{D}} L_{jk} L_{ki}\right]$

Prediction Step:

$$\boldsymbol{\mu}^{\boldsymbol{H}}_{t|t-1} = \begin{bmatrix} \mathbf{A}\boldsymbol{\mu}_{t-1|t-1} \\ \mathbf{0} \end{bmatrix}_{t|t-1}, \quad \boldsymbol{\Sigma}^{\boldsymbol{H}}_{t|t-1} = \begin{bmatrix} \mathbf{A}\boldsymbol{\Sigma}_{t-1|t-1}\mathbf{A}^\intercal + \mathbf{Q} & \boldsymbol{\Sigma}^{\boldsymbol{XW}} \\ (\boldsymbol{\Sigma}^{\boldsymbol{XW}})^\intercal & \boldsymbol{\Sigma}^{\boldsymbol{W}} \end{bmatrix}_{t|t-1},$$

$\boldsymbol{\mu}_{\boldsymbol{Y}} = \mathbf{C}\boldsymbol{\mu}_{t|t-1}, \quad \boldsymbol{\Sigma}_{\boldsymbol{Y}} = \mathbf{C}\boldsymbol{\Sigma}_{t|t-1}\mathbf{C}^\intercal + \mathbf{R}, \quad \boldsymbol{\Sigma}_{\boldsymbol{HY}} = \boldsymbol{\Sigma}^{\boldsymbol{H}}_{t|t-1}\mathbf{F}_t^\intercal$, where $\mathbf{F} = [\mathbf{C} \;\; \mathbf{0}]$

1^{st} **Update Step:**

$$\boldsymbol{\mu}^{\boldsymbol{H}}_{t|t} = \boldsymbol{\mu}^{\boldsymbol{H}}_{t|t-1} + \boldsymbol{\Sigma}_{\boldsymbol{HY}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}_{\boldsymbol{Y}}), \quad \boldsymbol{\Sigma}^{\boldsymbol{H}}_{t|t} = \boldsymbol{\Sigma}^{\boldsymbol{H}}_{t|t-1} - \boldsymbol{\Sigma}_{\boldsymbol{HY}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{HY}}^\intercal$$

Obtain the posterior PDF, $f(\boldsymbol{w}^{\boldsymbol{p}}_t|\boldsymbol{y}_{1:t}) = \mathcal{N}(\boldsymbol{w}^{\boldsymbol{p}}_t; \boldsymbol{\mu}^{\boldsymbol{W^p}}_{t|t}, \boldsymbol{\Sigma}^{\boldsymbol{W^p}}_{t|t})$

2^{nd} **Update Step:**

$$\boldsymbol{\mu}^{\overline{\boldsymbol{W^p}}}_{t|t} = \boldsymbol{\mu}^{\overline{\boldsymbol{W^p}}}_{t|t-1} + \mathbf{K}_t(\boldsymbol{\mu}^{\boldsymbol{W^p}}_{t|t} - \boldsymbol{\mu}^{\boldsymbol{W^p}}_{t|t-1}), \quad \boldsymbol{\Sigma}^{\overline{\boldsymbol{W^p}}}_{t|t} = \boldsymbol{\Sigma}^{\overline{\boldsymbol{W^p}}}_{t|t-1} + \mathbf{K}_t(\boldsymbol{\Sigma}^{\boldsymbol{W^p}}_{t|t} - \boldsymbol{\Sigma}^{\boldsymbol{W^p}}_{t|t-1})\mathbf{K}_t^\intercal,$$

$$\mathbf{K}_t = \boldsymbol{\Sigma}^{\boldsymbol{W^p}\overline{\boldsymbol{W^p}}}_{t|t-1}(\boldsymbol{\Sigma}^{\boldsymbol{W^p}}_{t|t-1})^{-1}, \quad \boldsymbol{\Sigma}^{\boldsymbol{W^p}\overline{\boldsymbol{W^p}}}_{t|t-1} = \boldsymbol{\Sigma}^{\overline{\boldsymbol{W^p}}}_{t|t-1}$$

Posterior moments of $\overrightarrow{\boldsymbol{LW}}$:

$$\boldsymbol{\mu}^{\overrightarrow{\boldsymbol{LW}}}_{t|t} = \boldsymbol{\mu}^{\overrightarrow{\boldsymbol{LW}}}_{t|t-1} + \mathbf{K}^{\mathbf{L}}_t(\boldsymbol{\mu}^{\overline{\boldsymbol{W^p}}}_{t|t} - \boldsymbol{\mu}^{\overline{\boldsymbol{W^p}}}_{t|t-1}), \quad \boldsymbol{\Sigma}^{\overrightarrow{\boldsymbol{LW}}}_{t|t} = \boldsymbol{\Sigma}^{\overrightarrow{\boldsymbol{LW}}}_{t|t-1} + \mathbf{K}^{\mathbf{L}}_t(\boldsymbol{\Sigma}^{\overline{\boldsymbol{W^p}}}_{t|t} - \boldsymbol{\Sigma}^{\overline{\boldsymbol{W^p}}}_{t|t-1})(\mathbf{K}^{\mathbf{L}}_t)^\intercal,$$

$$\mathbf{K}^{\mathbf{L}}_t = \boldsymbol{\Sigma}^{\overrightarrow{\boldsymbol{LW}}\overline{\boldsymbol{W^p}}}_{t|t-1}(\boldsymbol{\Sigma}^{\overline{\boldsymbol{W^p}}}_{t|t-1})^{-1}$$

return $\boldsymbol{\mu}_{t|t}$, $\boldsymbol{\Sigma}_{t|t}$, $\boldsymbol{\mu}^{\overrightarrow{\boldsymbol{LW}}}_{t|t}$, and $\boldsymbol{\Sigma}^{\overrightarrow{\boldsymbol{LW}}}_{t|t}$

E Additional Results for Case Study 2

E.1 Online inference of the variance and covariance terms in the full Q matrix

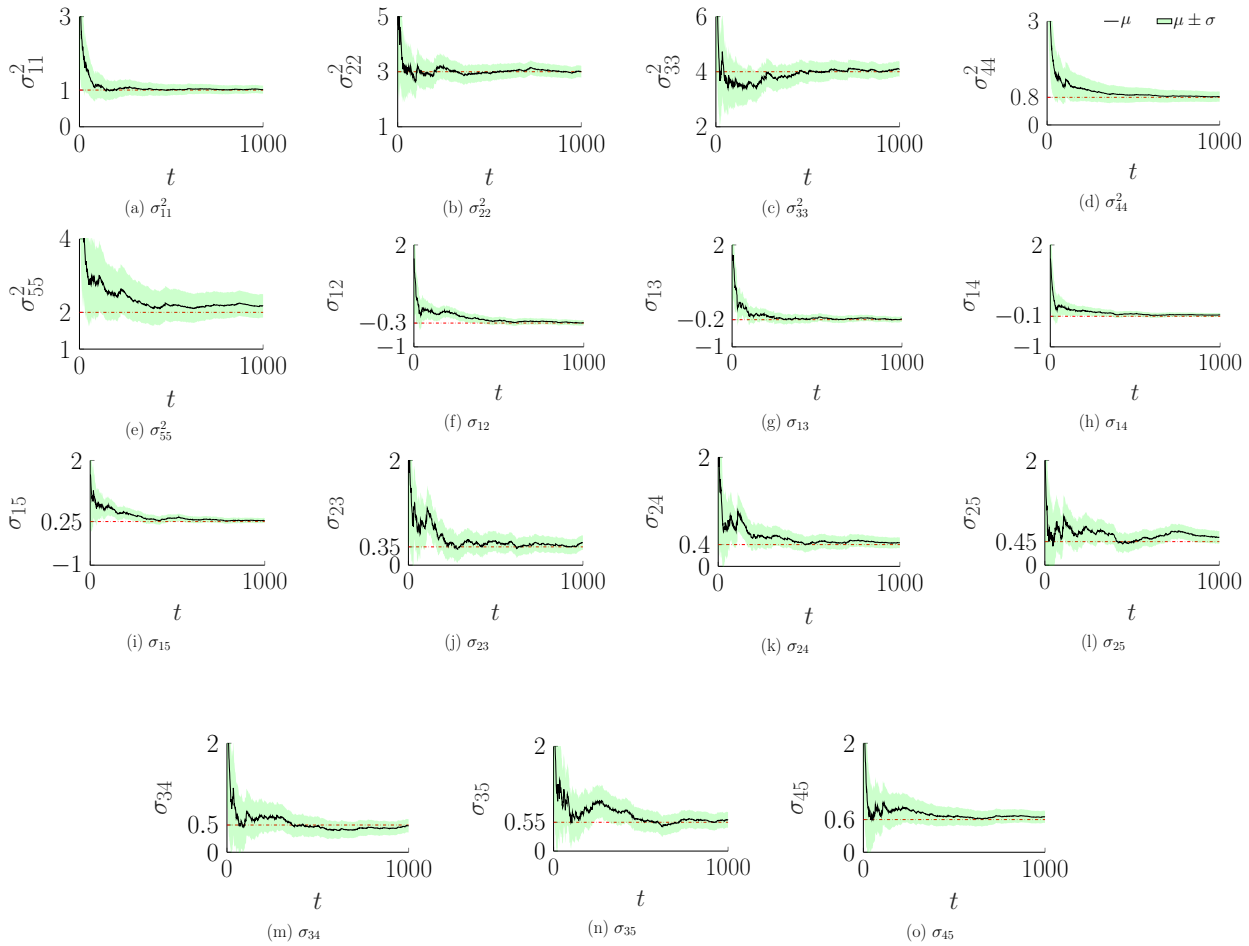


Figure 7: Online estimation of the error variance term and the covariance terms from the full **Q** matrix compared to their true values marked by the dashed red line. The estimated values are shown by the black solid line and their $\pm 1\sigma$ uncertainty bound is shown using the green shaded region.