14th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP14
Dublin, Ireland, July 9-13, 2023

Online Aleatory Uncertainty Quantification for Probabilistic Time Series Models

Bhargob Deka
Postdoctoral Researcher, Dept. of Civil, Geological, and Mining Engineering, Polytechnique Montréal, Montréal, Canada

James-A. Goulet
Professor, Dept. of Civil, Geological, and Mining Engineering, Polytechnique Montréal, Montréal, Canada

ABSTRACT: State-space models are a type of probabilistic approach used for time series forecasting. Such models involve unknown parameters for modeling both epistemic and aleatory uncertainties. Estimating the mean and epistemic uncertainty for the model can be computationally cheap. However, quantifying the model's aleatory uncertainty is typically the most computationally demanding task in the state estimation procedure. This article presents a new analytical Bayesian inference method called the approximate Gaussian variance inference (AGVI) that allows for online closed-form estimation of the error variance parameters that are used to quantify aleatory uncertainties. The case studies show that the AGVI method can exceed the performance of existing approaches in terms of its predictive capacity while being up to orders of magnitude faster.

1. Introduction

Engineering problems rely on probabilistic models for decision making tasks. One such approach are the state-space models that are often used for time series forecasting (Goulet, 2020; Murphy, 2012). Such models involve unknown parameters not only for modeling physical phenomena but also for quantifying the model's epistemic and aleatory uncertainties. In practice, estimating some of these parameters can be orders of magnitude more computationally demanding than others, and that prevents existing models from being scaled up for large-scale implementation (Sarkka and Nummenmaa, 2009).

For instance, estimating the mean and the epistemic uncertainty for the hidden state variables is computationally cheap as an analytical formulation exist for performing the Bayesian inference. On the other hand, the aleatory uncertainties are quantified by variance parameters in the process and observation error covariance matrices, which need to be known accurately for an exact state estimation (Mehra, 1970; Duník et al., 2017). In practice, obtaining optimal estimates for these unknown error variance parameters is typically the most computationally demanding task in the state estimation procedure. Even though the observation error matrix can be, in many situations, considered to be known from the measuring instrument specifications, it still remains a challenge to develop a computationally efficient method which is able to perform closed-form online estimation of the process error covariance matrix $\mathbf{Q}$ for multiple time series (Huang et al., 2020).

In this paper, we present the approximate Gaussian variance inference (AGVI) method for performing online inference of the error variance terms in the full $\mathbf{Q}$ matrix. The layout of this paper is as follows: Section 2 presents the problem statement,

Section 3 presents the methodology for the AGVI method, Section 4 presents two applied examples showing the application of the method where the first case study compares the performance of AGVI with the existing adaptive Kalman filtering (AKF) approaches (Mehra, 1970) and the second case study shows its application on real datasets from a concrete gravity dam. Finally, Section 5 presents the conclusion drawn from the results obtained using the AGVI method.

2. Problem Statement

For $\mathtt{D}$ time series, the global hidden state vector at time t is represented by $\boldsymbol{x}_t = [\boldsymbol{x}_t^1\ \boldsymbol{x}_t^2\ \cdots\ \boldsymbol{x}_t^j\ \cdots\ \boldsymbol{x}_t^{\mathtt{D}}]^\intercal$, where $\boldsymbol{x}_t^j,\ \forall j \in \{1,2,\cdots,\mathtt{D}\}$ refers to the state vector for the j^{th} time series. Similarly, the global transition matrix $\mathbf{A}$, the observation matrix $\mathbf{C}$, the process error covariance matrix $\mathbf{Q}$, and the observation error covariance matrix $\mathbf{R}$ are assembled block diagonally shown by

$$\begin{aligned}\mathbf{A} &= \text{blkdiag}[\mathbf{A}^1, \mathbf{A}^2, \cdots, \mathbf{A}^{\mathtt{D}}],\\ \mathbf{C} &= \text{blkdiag}[\mathbf{C}^1, \mathbf{C}^2, \cdots, \mathbf{C}^{\mathtt{D}}],\\ \mathbf{Q} &= \text{blkdiag}[\mathbf{Q}^1, \mathbf{Q}^2, \cdots, \mathbf{Q}^{\mathtt{D}}],\\ \mathbf{R} &= \text{blkdiag}[\mathbf{R}^1, \mathbf{R}^2, \cdots, \mathbf{R}^{\mathtt{D}}].\end{aligned}$$

The vector of process errors associated with the $\mathtt{D}$ time series is given by $\boldsymbol{w}_t = [\boldsymbol{w}_t^1\ \boldsymbol{w}_t^2\ \cdots\ \boldsymbol{w}_t^j\ \cdots\ \boldsymbol{w}_t^{\mathtt{D}}]^\intercal$, where $\boldsymbol{w}_t^j$ is the process error vector for the j^{th} time series. Moreover, each of the process error w in $\boldsymbol{w}_t^j$ is modeled as a zero-mean Gaussian random variable such that $w : W \sim \mathcal{N}(w; 0, \sigma_W^2)$. The global $\mathbf{Q}$ matrix associated with the vector $\boldsymbol{w}_t$ can be presented as follows

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

where $\mathbf{Q}^{i,j} \equiv \text{cov}(\boldsymbol{W}^i, \boldsymbol{W}^j)$ represents the covariance matrix between the two process errors $\boldsymbol{W}^i$ and $\boldsymbol{W}^j$ belonging to the i^{th} and j^{th} time series, respectively, where $i, j \in \{1,2,\cdots,\mathtt{D}\}$. Also, note that it is only possible to infer one error variance term σ_W^2 for any $\mathbf{Q}^i$ in the global $\mathbf{Q}$ matrix presented in Equation 1. This is due to the fact that only one unknown variable can be solved per equation. Hence, for multiple time series, the task is to infer the one error variance term associated with any i^{th} time series as well as the covariance term between any pair of i^{th} and j^{th} time series.

3. Methodology

This section presents the approximate Gaussian variance inference (AGVI) method for inferring the error variance and covariance terms associated with the multivariate process errors in case of multiple time series. At first, the core idea behind the AGVI method is presented through the univariate process error. Thereafter, the methodology is extended to the multivariate process errors.

3.1. Univariate Process Error

The main idea behind the AGVI method is explained through the univariate process error W. Given that W has a zero mean, it can be shown by definition that the expected value of the square of the process error W^2 is equal to the error variance parameter, i.e., $\mathbb{E}[W^2] = \sigma_W^2$. Considering that $W^2 \sim \mathcal{N}(w^2; \mathbb{E}[W^2], \text{var}(W^2))$ is Gaussian, the error variance parameter σ_W^2 is the same as the mean for the probability density function (PDF) of W^2, i.e., $\mathbb{E}[W^2]$. Now, assuming that this mean parameter is itself random, obtaining the posterior PDF of $\mathbb{E}[W^2]$ is the same as inferring the error variance term. In light of this fact, the AGVI method considers the process error variance term $\sigma_W^2 = \mathbb{E}[W^2]$ as a Gaussian random variable $\overline{W^2}$ such that

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

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

$$W \sim \mathcal{N}(w; 0, \overline{w^2}). \tag{3}$$

Further details regarding the proofs of these statements are provided by Deka (2022).

3.2. Multivariate Process Error

For multivariate process errors, let us consider the error vector $\boldsymbol{w} = \begin{bmatrix} w^1\ w^2\ \cdots\ w^i\ \cdots\ w^{\mathtt{D}} \end{bmatrix}^\intercal$, where $w^i,\ \forall i \in \{1,\ 2,\ \cdots, \mathtt{D}\}$ represents the one process error term for the i^{th} time series for which the error

variance term has to be inferred. Given that the process error random variable W is assumed to have a zero mean, the covariance term between i^{th} and j^{th} process error $\mathbf{Q}^{i,j}$ can be shown by

$$\text{cov}(W^i, W^j) \;=\; \mathbb{E}[W^i W^j]. \tag{4}$$

By leveraging Equation 4, we can formulate the covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ by

$$\boldsymbol{\Sigma}^{\boldsymbol{W}} = \begin{bmatrix} \mathbb{E}[(W^1)^2] & \mathbb{E}[W^1W^2] & \cdots & \mathbb{E}[W^1W^{\mathtt{D}}] \\ \vdots & \mathbb{E}[(W^2)^2] & \cdots & \mathbb{E}[W^2W^{\mathtt{D}}] \\ \vdots & \cdots & \ddots & \vdots \\ \text{sym.} & \cdots & \cdots & \mathbb{E}[(W^{\mathtt{D}})^2] \end{bmatrix}, \tag{5}$$

where $\text{var}(W^i) = \mathbb{E}[(W^i)^2]$ is the error variance for the i^{th} time series, and $\text{cov}(W^i, W^j) = \mathbb{E}[W^iW^j]$ is the covariance term between the two process errors for the i^{th} and j^{th} time series.

Let us consider that each of the product terms W^iW^j in Equation 5 is assumed to be a Gaussian random variable such that

$$W^iW^j \;\sim\; \mathcal{N}(w^iw^j; \mu^{W^iW^j}, (\sigma^{W^iW^j})^2), \tag{6}$$

where $\mathbb{E}[W^iW^j] = \mu^{W^iW^j}$ is the mean parameter and $\text{var}(W^iW^j) = (\sigma^{W^iW^j})^2$ is the variance. There are a total of $\mathtt{D}\cdot(\mathtt{D}+1)$ product terms that can be represented by the random vector $\boldsymbol{w^p} = [(w^1)^2\ (w^2)^2\ \cdots\ w^iw^j\ \cdots\ w^{\mathtt{D}}w^{\mathtt{D}\text{-}1}]^\intercal$ which is assumed to be Gaussian such that

$$\boldsymbol{W^p} \sim \mathcal{N}(\boldsymbol{w^p}; \boldsymbol{\mu}^{\boldsymbol{W^p}}, \boldsymbol{\Sigma}^{\boldsymbol{W^p}}), \tag{7}$$

where $\boldsymbol{\mu}^{\boldsymbol{W^p}}$ and $\boldsymbol{\Sigma}^{\boldsymbol{W^p}}$ represents the mean vector and the covariance matrix of $\boldsymbol{W^p}$. Similarly to the univariate case as shown in Section 3.1, the mean vector $\boldsymbol{\mu}^{\boldsymbol{W^p}}$ in Equation 7 is also assumed to be a Gaussian random variable represented by $\overline{\boldsymbol{W^p}}$ such that

$$\overline{\boldsymbol{W^p}} \sim \mathcal{N}(\overline{\boldsymbol{w^p}}; \boldsymbol{\mu}^{\overline{\boldsymbol{W^p}}}, \boldsymbol{\Sigma}^{\overline{\boldsymbol{W^p}}}). \tag{8}$$

where the random vector $\overline{\boldsymbol{w^p}}$ is

$$\overline{\boldsymbol{w^p}} = \left[\overline{(w^1)^2}\ \overline{(w^2)^2}\ \cdots\ \overline{(w^{\mathtt{D}})^2}\ \overline{w^1w^2}\ \cdots\ \overline{w^{\mathtt{D}\text{-}1}w^{\mathtt{D}}}\right]^\intercal.$$

The random variables in the vector $\overline{\boldsymbol{w^p}}$ are assumed to be independent of each other. The mean vector and covariance matrix for $\overline{\boldsymbol{W^p}}$ are

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

$$\boldsymbol{\Sigma}^{\overline{\boldsymbol{W^p}}} = \begin{bmatrix} (\sigma^{\overline{(W^1)^2}})^2 & 0 & \cdots & 0 \\ \vdots & (\sigma^{\overline{(W^2)^2}})^2 & \cdots & 0 \\ \vdots & \cdots & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ \text{sym.} & \cdots & \cdots & (\sigma^{\overline{W^{\mathtt{D}\text{-}1}W^{\mathtt{D}}}})^2 \end{bmatrix}. \tag{9}$$

Using the prior knowledge for $\overline{\boldsymbol{W^p}}$ defined in Equation 9, the first task is to obtain the prior predictive PDF of $\boldsymbol{W}^{\boldsymbol{p}}_{t|t-1}$, using which we can obtain the covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ defined in Equation 5 shown by

$$\boldsymbol{\Sigma}^{\boldsymbol{W}}_{t|t-1} = \begin{bmatrix} \mu^{\overline{(W^1)^2}} & \mu^{\overline{W^1W^2}} & \cdots & \mu^{\overline{W^1W^{\mathtt{D}}}} \\ \vdots & \mu^{\overline{(W^2)^2}} & \cdots & \mu^{\overline{W^2W^{\mathtt{D}}}} \\ \vdots & \cdots & \ddots & \vdots \\ \text{sym.} & \cdots & \cdots & \mu^{\overline{(W^{\mathtt{D}})^2}} \end{bmatrix}_{t|t-1}, \tag{10}$$

where the terms $\mathbb{E}[W^iW^j]$ in Equation 5 are substituted by the mean parameters of $\overline{\boldsymbol{W^p}}$, i.e., $\mu^{\overline{W^iW^j}}$. Moreover, we use the Cholesky factorization to ensure that $\boldsymbol{\Sigma}^{\boldsymbol{W}}_{t|t-1}$ shown by Equation 10 is positive semi-definite at any given time. This is carried out by constructing the prior knowledge of $\boldsymbol{\Sigma}^{\boldsymbol{W}}_{t|t-1}$ from an upper-triangular random matrix $\boldsymbol{L^W}$ where the vector of random variables in $\boldsymbol{L^W}$ is represented by $\overrightarrow{\boldsymbol{L^W}}$.

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

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

where the covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ defined in Equation 10 is obtained using the prior knowledge of $\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}$. The final task is to infer the covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ that needs two update steps. The first step involves the Gaussian conditional equations which are used to obtain the posterior PDF of $\boldsymbol{H}$ shown by

$$\boldsymbol{H}_{t|t} \sim \mathscr{N}(\boldsymbol{h}_t, \boldsymbol{\mu}^{\boldsymbol{H}}_{t|t}, \boldsymbol{\Sigma}^{\boldsymbol{H}}_{t|t}). \tag{11}$$

The second update step consists of several sub-steps. First, we use the posterior PDF $f(\boldsymbol{w}_t|\boldsymbol{y}_{1:t})$ obtained from Equation 11 and the Gaussian multiplicative approximation (GMA) equations (Goulet et al., 2021; Deka et al., 2021) to obtain the posterior PDF of $\boldsymbol{W}^{\boldsymbol{p}}_{t|t}$ such that $\boldsymbol{W}^{\boldsymbol{p}}_{t|t} \sim \mathscr{N}(\boldsymbol{w}^{\boldsymbol{p}}_t; \boldsymbol{\mu}^{\boldsymbol{W^p}}_{t|t}, \boldsymbol{\Sigma}^{\boldsymbol{W^p}}_{t|t})$. Second, we use the posterior PDF of $\boldsymbol{W^p}$ to update our knowledge for $\overline{\boldsymbol{W^p}}$ where the posterior mean, variance and covariance terms are

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

The last sub-step is to update the knowledge of $\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}$ using the posterior PDF of $\overline{\boldsymbol{W^p}}$ for which the moments are as follows

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

Both these steps are to be employed recursively with every new observation; first to obtain the covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{W}}$ and then to use it to update our knowledge for $\overline{\boldsymbol{W^p}}$. The detailed proofs for these statements are provided by Deka (2022).

4. Applied Examples

This section presents two case studies to demonstrate the application of AGVI for multiple time series. The first case study uses a simulated multivariate random walk model with a full process error covariance matrix $\mathbf{Q}$. The performance of the AGVI method is compared to the adaptive Kalman filter (AKF) methods, namely the *indirect correlation method* (ICM) (Mehra, 1970), the *adaptive limited memory filter* (ALMF) (Myers and Tapley, 1976), and the *sliding window variational adaptive Kalman filter* (SWVAKF) (Huang et al., 2020). Each of these AKF methods falls under separate categories; ICM is a correlation method, ALMF is a covariance-matching method (CMM), and SWVAKF is a variational Bayesian method. The second case study shows the application of AGVI on real displacement datasets obtained from a concrete dam in Québec, Canada.

4.1. *Case Study* 1 – *Multivariate Random Walk Model*

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

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

The state transition matrix $\mathbf{A}$ and the observation matrix $\mathbf{C}$ are defined as $\mathbf{A} = \boldsymbol{I}_3$, and $\mathbf{C} = \boldsymbol{I}_3$, The $\mathbf{Q}$ and the $\mathbf{R}$ matrices are defined as

$$\begin{aligned}
\mathbf{Q} &= \begin{bmatrix} 1 & -0.5 & -0.3 \\ -0.5 & 2 & 0.95 \\ -0.3 & 0.95 & 4 \end{bmatrix},\\
\mathbf{R} &= 10^{-4} \cdot \boldsymbol{I}_3,
\end{aligned}$$

where the off-diagonal covariance terms in the $\mathbf{Q}$ matrix are selected arbitrarily such that it is symmetric and positive-definite, i.e., the eigen values are positive. For AGVI, the prior knowledge for the augmented hidden states $\tilde{\boldsymbol{\mu}}_{0|0} = [\boldsymbol{\mu}_{0|0}; \boldsymbol{\mu}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}}_{0|0}]$ and $\tilde{\boldsymbol{\Sigma}}_{0|0} = \mathrm{blkdiag}(\boldsymbol{\Sigma}_{0|0}, \boldsymbol{\Sigma}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}}_{0|0})$ are initialized by

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_{0|0} &= [\mathbf{0}_3^{\intercal}\ 2\cdot\mathbf{1}_3^{\intercal}\ 0.8\cdot\mathbf{1}_3^{\intercal}]^{\intercal},\\
\tilde{\boldsymbol{\Sigma}}_{0|0} &= \mathrm{diag}([\mathbf{1}_3^{\intercal}\ 1\cdot\mathbf{1}_3^{\intercal}\ 0.5\cdot\mathbf{1}_3^{\intercal}]),
\end{aligned} \tag{12}$$

where $\mathbf{0}$ and $\mathbf{1}$ represent vector of zeros and ones, respectively. The mean vector and the covariance matrix for $\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}_{0|0} = [L_{11}\ L_{22}\ L_{33}\ L_{12}\ L_{13}\ L_{23}]^{\intercal}_{0|0}$ are

given by

$$\begin{aligned} \boldsymbol{\mu}_{0|0}^{\overrightarrow{L^{\boldsymbol{W}}}} &= [2 \cdot \mathbf{1}_3^\intercal \;\; 0.8 \cdot \mathbf{1}_3^\intercal]^\intercal, \\ \boldsymbol{\Sigma}_{0|0}^{\overrightarrow{L^{\boldsymbol{W}}}} &= \text{diag}([\,1 \cdot \mathbf{1}_3^\intercal \;\; 0.5 \cdot \mathbf{1}_3^\intercal]). \end{aligned} \tag{13}$$

For the AKF methods, the hidden states are initialized similarly to Equation 12, where the mean vector is $\boldsymbol{\mu}_{0|0} = \mathbf{0}_3$ and the covariance matrix is $\boldsymbol{\Sigma}_{0|0} = \boldsymbol{I}_3$. The hyperparameters for ICM include the stable Kalman gain ($\mathbf{K}$) and the auto-covariance lag parameter which are fixed to 0.99 and 1, whereas for ALMF, the initial $\mathbf{Q}$ matrix is chosen as $\boldsymbol{I}_3$. For SWVAKF, the same parameters are used as provided in the implementation code (Huang et al., 2020). Figure 1 compares the true values with the online hidden state estimates obtained using AGVI for a subset of elements from the $\mathbf{Q}$ matrix, namely σ_{22}^2, σ_{33}^2, σ_{12}, and σ_{13}. The true values for each element is shown by the dashed red line and the estimated values are shown by the black solid line and their $\pm 1\sigma$ uncertainty bounds are shown using the green shaded region. Table 1 shows the average RMSE values over five independent runs for

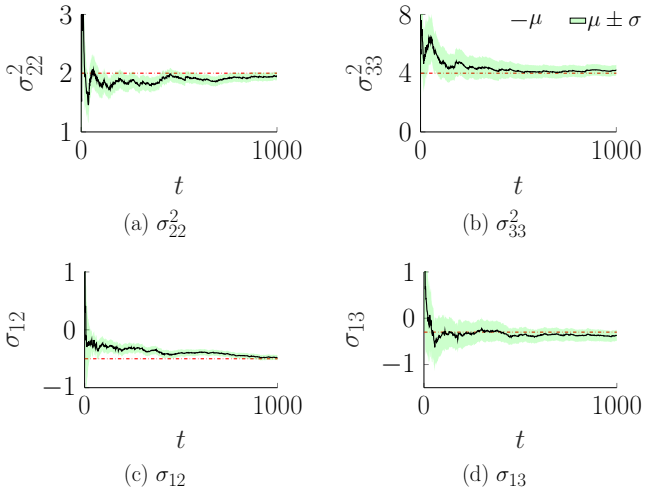


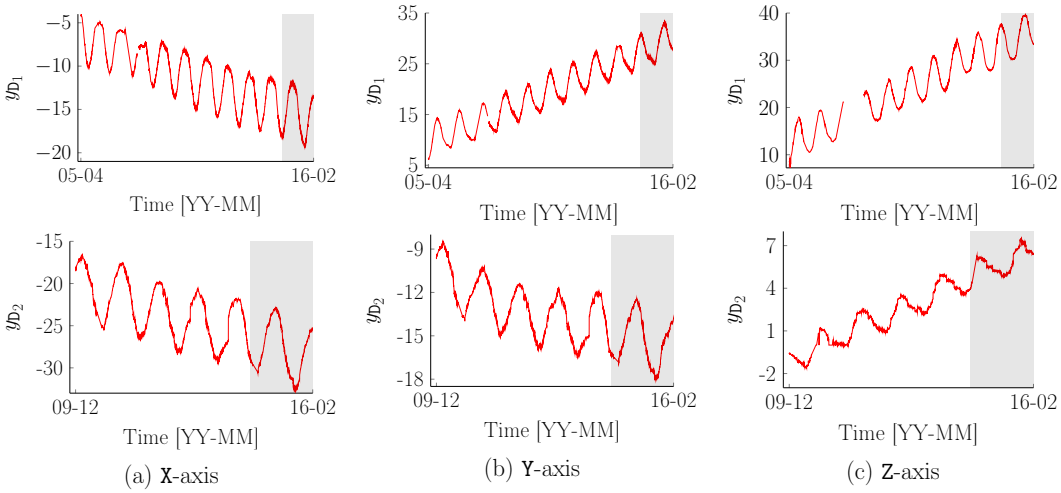
Figure 1: Online estimation of the error variance terms (a) σ_{22}^2, and (b) σ_{33}^2, and the covariance terms (c) σ_{12}, and (d) σ_{13} from the full $\mathbf{Q}$ matrix compared to their true values marked by the dashed red line. The estimated values are shown by the black solid line and their $\pm 1\sigma$ uncertainty bounds are shown using the green shaded region.

Table 1: Comparison of the average RMSE values for a subset of elements from the $\mathbf{Q}$ matrix as well as the computational time (s) for each method. The results are averaged over five independent runs. Each of the methods are picked from different AKF categories where AGVI and SWVAKF are Bayesian methods whereas ALMF is a covariance-matching method (CMM) and ICM is a correlation method. The variance terms and the covariance terms are represented by σ_{ii}^2 and σ_{ij}^2, $\forall i, j \in 1, \cdots, \mathtt{D}$.

Category	Methods	RMSE			Time (s)
		σ_{22}^2	σ_{33}^2	σ_{13}	
Bayesian	AGVI	**0.156**	**0.336**	**0.153**	0.642
Bayesian	SWVAKF	0.170	0.420	0.239	9.426
CMM	ALMF	0.206	0.750	0.161	0.056
Correlation	ICM	0.217	0.446	0.189	**0.003**

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 all methods in terms of predictive capacity for most of the variance and covariance terms. In comparison to SWVAKF which is also both a Bayesian and an online estimation method, it is more than an order of magnitude faster. The offline methods, i.e., ALMF and ICM, are faster compared to the Bayesian online methods, but can only provide point estimates which inhibits these methods in learning sequentially from data.

Furthermore, the normalized estimation error square (NEES) metric is used to check for statistical consistency (Bar-Shalom et al., 2004). The experiments conducted shows that on average ≈ 54 points lie outside the 95% C.I. that is comparable to the theoretical value of 50. Hence, the method provides statistically consistent estimates for the error variance terms. Table 4 in Appendix A provides the average number of points outside the 95% C.I. for the different prior initialization. The results presented here can be reproduced from the public code provided by Deka and Goulet (2023).



(a) X-axis (b) Y-axis (c) Z-axis

Figure 2: Plots showing the displacement datasets along X, Y, *and* Z*-axis collected by two sensors from a concrete dam in Canada.*

4.2. Case Study 2 – Dam Displacement

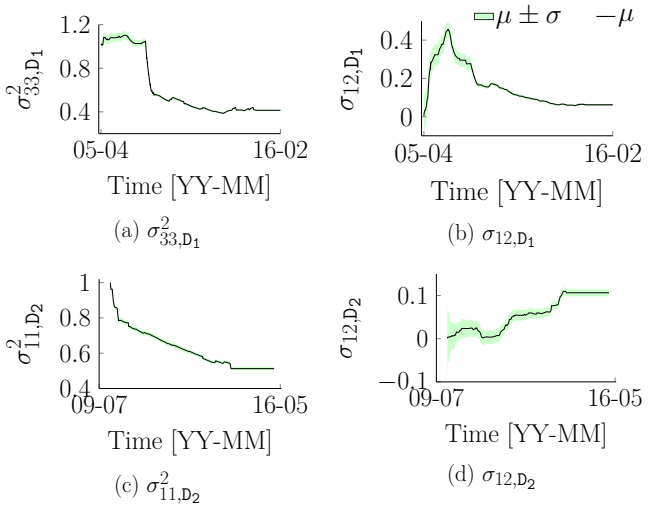
For this case study, two displacement datasets are used which are collected by two sensors from a concrete dam in Québec, Canada. Figure 2 shows the datasets along all three directions, i.e. X, Y, and Z-axis. The displacement datasets $\boldsymbol{y}_{\mathtt{D}_1}$ are available from April 2005 to February 2016 with a total of 9995 points, whereas the datasets $\boldsymbol{y}_{\mathtt{D}_2}$ are available from December 2009 to February 2016 with a total of 5667 points. A test set consisting of 1095 points is considered for all the datasets shown by the gray regions in Figure 2. Both datasets $\boldsymbol{y}_{\mathtt{D}_1}$ and $\boldsymbol{y}_{\mathtt{D}_2}$ are recorded with a non-uniform time-step size Δt. The most frequent time-step size is 12 hours for both datasets where the time-step size for $\boldsymbol{y}_{\mathtt{D}_1}$ varies in the range of 1 to 2792 hours, and for $\boldsymbol{y}_{\mathtt{D}_2}$ it varies in the range of 1 to 1032 hours.

The Bayesian dynamic linear model (BDLM) components used to model the patterns in the data are the *local trend* to model the baseline and the *kernel regression* component to model the periodic pattern. The process errors are modeled by a zero-mean vector and a full process error covariance matrix $\mathbf{Q}$ to be inferred using the multivariate AGVI method. For handling non-uniform time-step size, a reference time-step $\Delta t^{\mathtt{ref}}$ is chosen such that it is the most frequent time-step size in the datasets (Gaudot et al., 2019). The variance and the covariance parameters in the $\mathbf{Q}$ matrix are estimated for the reference time-step $\Delta t^{\mathtt{ref}}$, and for any time-step size Δt different than $\Delta t^{\mathtt{ref}}$, these parameter values are linearly scaled by the ratio between the current time-step and the reference time-step shown by

$$\mathbf{Q}^{\Delta t} = \mathbf{Q}^{\Delta t^{\mathtt{ref}}} \cdot \frac{\Delta t}{\Delta t^{\mathtt{ref}}},$$

where $\mathbf{Q}^{\Delta t}$ is the updated process error covariance matrix at the current time-step and $\mathbf{Q}^{\Delta t^{\mathtt{ref}}}$ is the covariance matrix for the reference time-step.

The prior knowledge for the hidden states are defined using the default values provided by the OpenBDLM library (Gaudot et al., 2019). The optimized values for the kernel length parameters obtained using the Newton-Raphson are provided in Appendix B. The observation error covariance matrix $\mathbf{R}$ is set to $10^{-6} \cdot \boldsymbol{I}_3$ such that the process errors model the residuals considering that the measurements from the sensors are exact. As there are three time series from each sensor, the $\mathbf{Q}$ matrix has a size of 3×3 with three variance terms and three covariance terms to be evaluated. The prior mean vector and the covariance matrix for $\overrightarrow{\boldsymbol{W}}_{0|0} = [L_{11}\ L_{22}\ L_{33}\ L_{12}\ L_{13}\ L_{23}]^{\mathsf{T}}_{0|0}$, are provided in Appendix B. Figure 3 shows the online estimation for one variance and covariance term, chosen arbitrarily from the full $\mathbf{Q}$ matrix for both datasets $\boldsymbol{y}_{\mathtt{D}_1}$ and $\boldsymbol{y}_{\mathtt{D}_2}$. The estimated values are shown by the



(a) $\sigma^2_{33,\mathtt{D}_1}$ (b) $\sigma_{12,\mathtt{D}_1}$ (c) $\sigma^2_{11,\mathtt{D}_2}$ (d) $\sigma_{12,\mathtt{D}_2}$

Figure 3: Online estimation of one error variance and covariance term chosen arbitrarily from the full $\mathbf{Q}$ *matrix for both datasets* $\boldsymbol{y}_{\mathtt{D}_1}$ *and* $\boldsymbol{y}_{\mathtt{D}_2}$*; where the terms are (a)* $\sigma^2_{33,\mathtt{D}_1}$*, (b)* $\sigma_{12,\mathtt{D}_1}$*, (c)* $\sigma^2_{11,\mathtt{D}_2}$*, and (d)* $\sigma_{12,\mathtt{D}_2}$*. The estimated values are shown by the black solid line and their* $\pm 1\sigma$ *uncertainty bounds are shown using the green shaded region.*

black solid line and their $\pm 1\sigma$ uncertainty bounds are shown using the green shaded region.

The predictive performance of using AGVI is compared with the one obtained using the Newton-Raphson method where the variance parameters are learned offline through optimization. Table 2 shows the test-set root mean square error (RMSE) and log-likelihood values obtained using AGVI and the Newton-Raphson method for the two datasets along all three axis. Table 3 compares the two methods in terms of computational time required, i.e., optimization time and training time expressed in seconds. The results show that AGVI has an accuracy comparable to the Newton-Raphson in terms of RMSE and outperforms it in terms of log-likelihood. Moreover, AGVI is orders of magnitude more computationally efficient than Newton-Raphson as it facilitates online learning of the process error variance and covariance terms, thereby avoiding the parameter optimization step. This example shows that AGVI is applicable to real case studies for evaluating the full $\mathbf{Q}$ matrix involving multiple time series.

Table 2: Root mean square error (RMSE) and log-likelihood values obtained with the AGVI and the Newton-Raphson methods for the displacements datasets $\boldsymbol{y}_{\mathtt{D}_1}$ *and* $\boldsymbol{y}_{\mathtt{D}_2}$ *along all three axis.*

	RMSE		Log-likelihood	
method	AGVI	Newton-Raphson	AGVI	Newton-Raphson
$\mathtt{X}_{\mathtt{D}_1}$	**0.598**	0.622	**−1576.8**	−1585.1
$\mathtt{Y}_{\mathtt{D}_1}$	**0.35**	**0.35**	**−556.76**	−630.38
$\mathtt{Z}_{\mathtt{D}_1}$	**0.64**	**0.64**	**−1031**	−1034.1
$\mathtt{X}_{\mathtt{D}_2}$	**0.57**	**0.57**	**−992.81**	−1069.2
$\mathtt{Y}_{\mathtt{D}_2}$	**0.84**	0.85	**−1414.80**	−2303.9
$\mathtt{Z}_{\mathtt{D}_2}$	**0.17**	**0.17**	**372.37**	361.43

Table 3: Comparison of optimization time (s) and training time (s) using AGVI and the Newton-Raphson method.

	Optimization Time (s)		Training Time (s)	
method	AGVI	Newton-Raphson	AGVI	Newton-Raphson
$\boldsymbol{y}_{\mathtt{D}_1}$	0	2894	61	51
$\boldsymbol{y}_{\mathtt{D}_2}$	0	961	34	26

5. Conclusion

This paper presents an analytical Bayesian method for performing closed-form inference of the error variance and covariance terms in the full $\mathbf{Q}$ matrix. The case study 1 shows the application of AGVI for a multivariate random walk model with a full $\mathbf{Q}$ matrix using synthetic data and compares its performance with the adaptive Kalman filtering (AKF) methods in the literature. The results show that AGVI outperforms all methods in terms of predictive capacity for most of the variance and covariance terms. In comparison to SWVAKF which is both a Bayesian and an online estimation method, it is more than an order of magnitude faster. The case study 2 shows the application of AGVI on real displacement datasets from a concrete dam in Canada. In comparison with the Newton-Raphson method, AGVI has a comparable accuracy in terms of RMSE while outperforming it in terms of log-likelihood. However, the online learning capacity of AGVI makes it computationally efficient com-

pared to Newton-Raphson which is an offline optimization method. Hence, the AGVI method has the capacity for online estimation of aleatory uncertainties in state-space models.

6. REFERENCES

Bar-Shalom, Y., Li, X. R., and Kirubarajan, T. (2004). *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons.

Deka, B. (2022). "Analytical Bayesian parameter inference for probabilistic models with engineering applications." Ph.D. thesis, Polytechnique Montréal, Polytechnique Montréal.

Deka, B. and Goulet, J.-A. (2023). "AGVI." *GitHub repository*. Accessed: 2023-01-20.

Deka, B., Ha Nguyen, L., Amiri, S., and Goulet, J.-A. (2021). "The Gaussian multiplicative approximation for state-space models." *Structural Control and Health Monitoring*, e2904.

Duník, J., Straka, O., Kost, O., and Havlík, J. (2017). "Noise covariance matrices in state-space models: A survey and comparison of estimation methods—part I." *International Journal of Adaptive Control and Signal Processing*, 31(11), 1505–1543.

Gaudot, I., Nguyen, L., S., K., and Goulet, J.-A. (2019). "OpenBDLM, an Open-Source Software for Structural Health Monitoring using Bayesian Dynamic Linear Models." *GitHub repository*. Accessed: 2022-09-22.

Goulet, J.-A. (2020). *Probabilistic machine learning for civil engineers*. MIT Press.

Goulet, J.-A., Nguyen, L. H., and Amiri, S. (2021). "Tractable approximate Gaussian inference for Bayesian neural networks." *Journal of Machine Learning Research*, 22(251), 1–23.

Huang, Y., Zhu, F., Jia, G., and Zhang, Y. (2020). "A slide window variational adaptive Kalman filter." *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67(12), 3552–3556.

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

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT Press.

Myers, K. and Tapley, B. (1976). "Adaptive sequential estimation with unknown noise statistics." *IEEE Transactions on Automatic Control*, 21(4), 520–523.

Sarkka, S. and Nummenmaa, A. (2009). "Recursive noise adaptive Kalman filtering by variational Bayesian approximations." *IEEE Transactions on Automatic control*, 54(3), 596–600.

A. CONSISTENCY CHECK IN CASE STUDY 1

Table 4: Average number of points outside the 95% *probability region for the different prior initialization of* $\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}_{0|0}}$ *such that* $\boldsymbol{\mu}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}}}_{0|0}=[\alpha\cdot\mathbf{1}_3^\intercal\ \beta\cdot\mathbf{1}_3^\intercal]^\intercal$ *while considering the same covariance matrix as defined in Equation 13. Given that the length of the time series is* 1000 *and a* 95% *confidence interval (C.I.), the number of acceptable points outside the bounds is* 50*. Each column presents the average value computed using five different runs for one combination of* $\{\alpha,\beta\}$.

	$\{1.6, 0.6\}$	$\{1.8, 0.8\}$	$\{2, 1\}$	Mean
NEES	50.8	52.4	58.4	53.93

B. PARAMETERS IN CASE STUDY 2

The kernel length parameters for the three datasets from each sensor are obtained by offline optimization using the Newton-Raphson method such that $\{\ell_{\mathtt{X}_{\mathtt{D}_1}}=0.350,\ \ell_{\mathtt{Y}_{\mathtt{D}_1}}=0.362, \ell_{\mathtt{Z}_{\mathtt{D}_1}}=0.347\}$ and $\{\ell_{\mathtt{X}_{\mathtt{D}_2}}=0.288,\ \ell_{\mathtt{Y}_{\mathtt{D}_2}}=0.289, \ell_{\mathtt{Z}_{\mathtt{D}_2}}=0.602\}$.

The prior mean vector and the covariance matrix for $\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}_{0|0}}=[L_{11}\ L_{22}\ L_{33}\ L_{12}\ L_{13}\ L_{23}]^\intercal_{0|0}$, are given by

$$\begin{aligned}
\boldsymbol{\mu}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}_{\mathtt{D}_1}}}_{0|0} &= [1\cdot\mathbf{1}_3\ \ \text{1e-03}\cdot\mathbf{1}_3]^\intercal,\\
\boldsymbol{\mu}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}_{\mathtt{D}_2}}}_{0|0} &= [1\ 1\ 0.5\ \ \text{1e-03}\cdot\mathbf{1}_3]^\intercal,\\
\boldsymbol{\Sigma}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}_{\mathtt{D}_1}}}_{0|0} &= \boldsymbol{\Sigma}^{\overrightarrow{\boldsymbol{L}^{\boldsymbol{W}}_{\mathtt{D}_2}}}_{0|0} = \text{diag}([\ \text{1e-04}\cdot\mathbf{1}_3\ \ \text{1e-02}\cdot\mathbf{1}_3]).
\end{aligned}$$