

Real-time Anomaly Detection with Bayesian Dynamic Linear Models

LUONG HA NGUYEN* and JAMES-A. GOULET

Department of Civil, Geologic and Mining Engineering
ECOLE POLYTECHNIQUE DE MONTREAL, CANADA

May 22, 2019

Abstract

A key goal in Structural Health Monitoring is to detect abnormal events in a structure's behavior by interpreting its observed responses over time. The goal is to develop an anomaly-detection method that (i) is robust towards false alarm, and (ii) capable of performing real-time analysis. The majority of anomaly detection approaches are currently operating over batches of data for which the model parameters are assumed to be constant over time, and to be equal to the values estimated during a fixed-size training-period. This assumption is not suited for the real-time anomaly detection where model parameters need to be treated as time-varying quantities. This paper presents how this issue is tackled by combining Rao-Blackwellized Particle Filtering (RBPF) with the Bayesian Dynamic Linear Models (BDLMs). The BDLMs, which is a special case of state space models, allow decomposing time-series into a vector of hidden state variables. The RBPF employs the sequential Monte Carlo method to learn model parameters continuously as the new observations are collected. The potential of the new approach is illustrated on the displacement data collected from a dam in Canada. The approach succeeds in detecting the anomaly caused by the refection work on the dam as well as the artificial anomalies that are introduced on the original dataset. The new method opens the way for monitoring the structure's health and conditions in real-time.

Keywords: Anomaly detection, Online learning, Structural Health Monitoring, Bayesian, Dynamic linear models, Particle filter

1 Introduction

Structural Health Monitoring (SHM) is a research field studying the state of civil infrastructure based on sensor responses such as displacements, elongations and accelerations [7, 24]. SHM attempts to gain insightful information about the state of a structure by interpreting its observed responses over time. The key role of SHM is to detect the changes in behavior of structures, that is, *anomalies*, in order to allow for preventive infrastructure maintenance in time. The improvements in sensor technologies allow civil infrastructure to be monitored

*Corresponding author: luongha.nguyen@gmail.com

continuously over time. There is a need for anomaly detection methodologies that are capable of performing real-time analysis, while being robust towards false alarms. *Real time* hereby, means performing the data analysis as the new observations become available.

For time-series data, the majority of existing anomaly detection approaches operates using a batch learning procedure [4, 36] in which the model parameters are assumed to be constant over time, and are estimated by minimizing a cost function, such as the prediction error, within a fixed training-period. Several approaches based on the batch learning procedure have been proposed for anomaly detection such as *Autoregressive Integrated Moving Average* [12], *Holt-Winters* [42], *Seasonal-Trend Decomposition* [8], *Long Short Term Memory* [22, 25], and among others. A common limitation of these approaches is that the assumption of invariant model parameters is no longer suitable for real-time anomaly detection where the underlying process in stream data can change over time [2, 21]. Another aspect is that these methods are computationally demanding for large datasets because when a new data point arrives, the entire model needs to be retrained in order to estimate the new model parameters. To address this limitation, Sejnowski and Rosenberg [41] has proposed a sliding window technique in which a small dataset from the past is kept for the learning purpose. This technique enables the batch-learning algorithm to learn continuously using only the data in the selected window and the new data point. However, the window length may become an issue because a short window-length might not have enough information for the learning purpose, while a long window-length can slow down the learning process. An *online learning* [10, 19, 26] algorithm is needed for estimating the model parameters of each new observation in order to rapidly adapt to changes. In the field of machine learning, examples of the online learning are *Extreme Learning Machine* [20], *Bayesian Online Changepoint Detection* [1], and *Random Cut Forest* [17]. In addition to these methods, *Hierarchical Temporal Memory* mimicking the architecture and processes of cortical neurons shows potential for the real-time anomaly detection [3].

In the field of civil engineering, several studies have adopted machine learning methods such as *Particle-filter-based model* [6, 9] for the online learning purpose. Existing applications of such methods typically require specific information about the structure which is not suited for a widespread deployment across thousands of bridges and dams that are all different from one to another. More recently, Nguyen and Goulet [32] have proposed an anomaly detection method based on the theory of *Bayesian Dynamic Linear Models* (BDLMs) that decomposes the observed responses into a vector of the hidden state variables. In order to detect anomalies, this approach considers the prior probability of anomalies, the anomaly's kinematic model and the probability to transition from a normal to an abnormal state. This approach is a promising path towards a deployment on large-scale applications for civil infrastructure. The challenge remaining to be addressed is that the approach currently operates using a batch learning procedure.

This paper proposes a new approach combining the existing BDLMs with the theory of *Rao-Blackwellized Particle Filter* (RBPF) [11]. The RBPF is a variant of the *Particle Filter* [28] that approximates the posterior *Probability Density Function* (PDF) for some hidden state variables using the importance-sampling method. The main idea of RBPF is to estimate analytically the posterior PDF for hidden state variables using Kalman filter equations [28] and approximate the posterior PDF for the model parameters using sampling. The RBPF has been applied to problems such as robot localization [15, 16], visual objection

tracking [40], online parameter estimation [34, 35], and anomaly detection for environmental data [18]. The contribution of this paper is to enable the existing BDLMs (i) to perform anomaly detection in real-time and (ii) to provide the real-time estimation of the hidden state variables as well as the model parameters.

The paper is organized as follows. The section 2 presents the state estimation theory for the BDLMs. Section 3 describes the mathematical formulation for estimating the model parameters. Section 4 presents the application of the new approach to the displacement data collected on a dam in Canada. Section 5 discusses about the advantages as well as the drawbacks of the proposed approach.

2 Bayesian Dynamic Linear Models

The section provides an overview of the theory of the *Bayesian Dynamic Linear Models* (BDLMs).

2.1 Hidden State Estimation

BDLMs are a special case of *State Space Models* which are described by a transition model and an observation model. The transition model describes the time-dependent relationships between the hidden state variables so that

$$\mathbf{x}_t = \mathbf{A}_t \mathbf{x}_{t-1} + \mathbf{w}_t, \quad (1)$$

where \mathbf{x}_t is the hidden state variable, \mathbf{A}_t is the transition matrix, and \mathbf{w}_t is the model error. The model error is assumed to be a multivariate Gaussian distribution with mean zero and covariance matrix \mathbf{Q}_t i.e. $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t)$. The dependence of the observations on the hidden state variables is described by the observation model

$$\mathbf{y}_t = \mathbf{C}_t \mathbf{x}_t + \mathbf{v}_t, \quad (2)$$

where \mathbf{y}_t is the observation, \mathbf{C}_t is the observation matrix, and \mathbf{v}_t is the Gaussian observation noise with mean zero and covariance matrix \mathbf{R}_t , i.e. $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_t)$. The model matrices $\{\mathbf{A}_t, \mathbf{C}_t, \mathbf{Q}_t, \mathbf{R}_t\}$ involve a vector of unknown parameters $\boldsymbol{\theta}_t$ to be inferred from the data. The hidden state variables at time t in Equations 1 and 2, are estimated using the Kalman Filter (KF) algorithm [28]. The KF algorithm separates in two main steps: *predicition step* and *update step*. The prediction step performs the estimation of the prior state estimate $p(\mathbf{x}_t | \boldsymbol{\theta}_t, \mathbf{y}_{1:t-1})$ while the update step is employed to estimate the posterior state estimate $p(\mathbf{x}_t | \boldsymbol{\theta}_t, \mathbf{y}_{1:t})$. The prior and posterior estimate states at time t are described by the multivariate Gaussian distribution following

$$\begin{aligned} p(\mathbf{x}_t | \boldsymbol{\theta}_t, \mathbf{y}_{1:t-1}) &= \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}) \\ p(\mathbf{x}_t | \boldsymbol{\theta}_t, \mathbf{y}_{1:t}) &= \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}), \end{aligned} \quad (3)$$

where $\boldsymbol{\mu}_{t|t-1}$ is the prior expected value, $\boldsymbol{\Sigma}_{t|t-1}$ is the prior covariance matrix, $\boldsymbol{\mu}_{t|t}$ is the posterior expected value, and $\boldsymbol{\Sigma}_{t|t}$ is the posterior covariance matrix. The KF algorithm is summarized in its short form as

$$(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}, \mathcal{L}_t) = \text{Filter}(\boldsymbol{\mu}_{t-1|t-1}, \boldsymbol{\Sigma}_{t-1|t-1}, \mathbf{y}_t, \mathbf{A}_t, \mathbf{C}_t, \mathbf{Q}_t, \mathbf{R}_t), \quad (4)$$

where \mathcal{L}_t is the marginal likelihood at time t . The full details for \mathcal{L}_t is presented in Section 2.2.

BDLMs decompose the observed structural responses into a vector of hidden state variables. These hidden state variables can be a *baseline* component to model the structural behavior, a *local trend* component to describe the rate of change in the baseline component, a *periodic* component to describe environmental conditions, and an *autoregressive* component to capture time-dependent model errors. If changes occur in the local trend component, a *local acceleration* component is employed to model its rate of change. Further details regarding the generic components and their hidden state variables are provided by Goulet [14] and West & Harrison [43].

2.2 Model Parameter Estimation

The typical primary approach for optimizing model parameters is the *Maximum Likelihood Estimation* (MLE). The optimal vector of model parameters $\boldsymbol{\theta}_t^*$ is obtained by maximizing the joint prior probability of observations with the assumption that the observations $\mathbf{y}_{1:T}$ are independent of each other so that

$$p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_t) = \prod_{t=1}^T \underbrace{p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}_t)}_{\mathcal{L}_t}, \quad (5)$$

where $p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}_t)$, denoted as \mathcal{L}_t , is the marginal likelihood at time t . With BDLMs, the marginal likelihood is a multivariate Gaussian distribution following

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}_t) = \mathcal{N}(\mathbf{y}_t; \mathbf{C}_t(\boldsymbol{\theta}_t)\boldsymbol{\mu}_{t|t-1}, \mathbf{C}_t(\boldsymbol{\theta}_t)\Sigma_{t|t-1}\mathbf{C}_t(\boldsymbol{\theta}_t)^T + \mathbf{R}_t(\boldsymbol{\theta}_t)). \quad (6)$$

In order to avoid the zero-underflow issue, the prior probability of observation is transformed in the natural logarithm space, so that Equation 5 is rewritten as

$$\ln p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_t) = \sum_{t=1}^T \ln p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}_t). \quad (7)$$

The optimal vector of model parameters $\boldsymbol{\theta}_t^*$ is learned during a specific training-set using the *Newton-Raphson* (NR) algorithm [13]. This procedure is called *batch learning* in which, $\boldsymbol{\theta}_t$ is assumed to be independent of time. The performance of the NR algorithm is dependent upon the initial values for the model parameters. Poor guesses for the initial values is prone to lead to a local maximum.

2.3 Anomaly Detection

The current anomaly-detection methodology in the BDLMs is using the *Switching Kalman Filter* (SKF) [27]. The SKF enables BDLMs to model different states (e.g., normal or abnormal) of a structure over time steps. Each model class has its own model matrices $\{\mathbf{A}_t, \mathbf{C}_t, \mathbf{Q}_t, \mathbf{R}_t\}$, and state probability at each time step. Note that the normal and abnormal states correspond to the stationary and non-stationary regimes, respectively. The SKF algorithm for the BDLMs includes a *filter* step and a *collapse* step.

SKF-Filter step

The SKF-Filter step employs the same Kalman equations as presented in Equation 4. Because the *Markov-switching variable* $s_t \in \{1, 2, \dots, S\}$ is employed in the SKF, the notation for the Kalman equations described in Equation 4 must be adapted. Note that each Markov-switching variable is related to a *filtering model* defined by its model matrices. The Markov-switching variables at time t and $t - 1$ are denoted s_t and s_{t-1} , respectively. The Kalman filter algorithm described above is summarized in its short form as

$$(\boldsymbol{\mu}_{t|t}^{i(j)}, \boldsymbol{\Sigma}_{t|t}^{i(j)}, \mathcal{L}_t^{i(j)}) = \text{Filter}(\boldsymbol{\mu}_{t-1|t-1}^j, \boldsymbol{\Sigma}_{t-1|t-1}^i, \mathbf{y}_t, \mathbf{A}_t^j, \mathbf{C}_t^j, \mathbf{Q}_t^{i(j)}, \mathbf{R}_t^j) \quad (8)$$

where the *superscript* inside the parentheses $i(j)$ refers to the current state j at the time t given the state i at time $t - 1$.

SKF-Collapse step

In the SKF-collapse step, the mean vector $\boldsymbol{\mu}_{t|t}^j$ and covariance matrix $\boldsymbol{\Sigma}_{t|t}^j$ for the Markov-switching variables j at time t , are evaluated by collapsing the Gaussian mixture resulting from Equation 8 into a single Gaussian density [28]. The short-form notation for the collapse step is

$$(\boldsymbol{\mu}_{t|t}^j, \boldsymbol{\Sigma}_{t|t}^j, \pi_{t|t}^j) = \text{Collapse}(\boldsymbol{\mu}_{t|t}^{i(j)}, \boldsymbol{\Sigma}_{t|t}^{i(j)}, \mathbf{W}_{t-1|t}^{i(j)}), \quad (9)$$

where $\mathbf{W}_{t-1|t}^{i(j)}$ is the state switching probability i.e. $p(s_{t-1} = i | s_t = j, \mathbf{y}_{1:t})$. Because of the presence of the Markov-switching variables, the marginal log-likelihood is defined as

$$\begin{aligned} \ln p(\mathbf{y}_{1:T} | \boldsymbol{\theta}_t) &= \sum_{t=1}^T \underbrace{\ln p(\mathbf{y}_t | \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_t)}_{\mathcal{L}_t} \\ &= \sum_{t=1}^T \ln \left[\sum_{j=1}^S \sum_{i=1}^S \underbrace{\mathcal{L}^{i(j)} \cdot p(s_t = j | s_{t-1} = i)}_{\mathbf{Z}^{i(j)}} \underbrace{p(s_{t-1} = i | \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_t)}_{\pi_{t-1|t-1}^i} \right], \end{aligned} \quad (10)$$

where $\mathbf{Z}^{i(j)}$ is the transition probability and $\pi_{t-1|t-1}^i$ is the previous state probability. The vector of optimal model parameters, $\boldsymbol{\theta}_t^*$, is estimated using the MLE as presented in Section 2.2. The full description for the anomaly detection methodology is provided by Nguyen and Goulet [32]. For the simplicity purpose, SKF-filter and SKF-collapse steps are summarized in a short form as

$$(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}, \mathcal{L}_t, \pi_{t|t}) = \text{SKF}(\boldsymbol{\mu}_{t-1|t-1}, \boldsymbol{\Sigma}_{t-1|t-1}, \mathbf{y}_t, \mathbf{A}_t, \mathbf{C}_t, \mathbf{Q}_t, \mathbf{R}_t, \mathbf{Z}_t, \pi_{t-1|t-1}). \quad (11)$$

3 Rao-Blackwellized Particle Filter

This section presents the mathematical formulation for *Rao-Blackwellized Particle Filter* (RBPF) [11] and a framework architecture for the *online learning* procedure in BDLMs. The RBPF employs the analytical Kalman equations for estimating the posterior PDF of hidden states \mathbf{x}_t and the *Sequential Importance Sampling* (SIS) [39] to approximate the posterior PDF of model parameters, $\boldsymbol{\theta}_t$. This approach allows learning continuously both the hidden state variables and the model parameters as new data points are collected.

3.1 Mathematical Formulation

The posterior PDF for both the hidden state variables and the model parameters, is theoretically defined as

$$p(\mathbf{x}_{1:t}, \boldsymbol{\theta}_{0:t} | \mathbf{y}_{1:t}) = p(\mathbf{x}_{1:t} | \boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t}) \cdot p(\boldsymbol{\theta}_{0:t} | \mathbf{y}_{1:t}), \quad (12)$$

where $p(\mathbf{x}_{1:t} | \boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t})$ is evaluated using the filtering equations presented in Section 2.1, and $p(\boldsymbol{\theta}_{0:t} | \mathbf{y}_{1:k})$ is obtained using SIS with a set of particles. According to Bayes theorem, the posterior PDF for model parameters can be written as follows

$$\begin{aligned} p(\boldsymbol{\theta}_{0:t} | \mathbf{y}_{1:t}) &\propto p(\mathbf{y}_t | \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}) \cdot p(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{0:t-1}, \mathbf{y}_{1:t-1}) \cdot p(\boldsymbol{\theta}_{0:t-1} | \mathbf{y}_{1:t-1}) \\ &\propto p(\mathbf{y}_t | \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_t) \cdot p(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}) \cdot p(\boldsymbol{\theta}_{0:t-1} | \mathbf{y}_{1:t-1}), \end{aligned} \quad (13)$$

where the second formula takes advantage of Markov's assumptions for the transition prior $p(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})$, and $p(\boldsymbol{\theta}_{0:t-1} | \mathbf{y}_{1:t-1})$ is the posterior from the previous time step. Equation 13 is written as being proportional to the posterior because the normalization constant is analytically intractable. The PDF for model parameters can be approximated using the importance sampling method. If the proposal distribution is chosen following

$$q(\boldsymbol{\theta}_{0:t} | \mathbf{y}_{1:t}) = q(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{0:t-1}, \mathbf{y}_{1:t}) \cdot q(\boldsymbol{\theta}_{0:t-1} | \mathbf{y}_{1:t-1}), \quad (14)$$

where the previous sets of particles $\boldsymbol{\theta}_{0:t-1}$ do not depend on future observations \mathbf{y}_t i.e. $q(\boldsymbol{\theta}_{0:t-1} | \mathbf{y}_{1:t-1}) \equiv q(\boldsymbol{\theta}_{0:t-1} | \mathbf{y}_{1:t})$. The importance weights can then be defined following

$$\begin{aligned} w_t &= \frac{p(\boldsymbol{\theta}_{0:t} | \mathbf{y}_{1:t})}{q(\boldsymbol{\theta}_{0:t} | \mathbf{y}_{1:t})} \\ &\propto \frac{p(\mathbf{y}_t | \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_t) \cdot p(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})}{q(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{0:t-1}, \mathbf{y}_{1:t})} \cdot \frac{p(\boldsymbol{\theta}_{0:t-1} | \mathbf{y}_{1:t-1})}{q(\boldsymbol{\theta}_{0:t-1} | \mathbf{y}_{1:t-1})} \\ &= \frac{p(\mathbf{y}_t | \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_t) \cdot p(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})}{q(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{0:t-1}, \mathbf{y}_{1:t})} \cdot w_{t-1}. \end{aligned} \quad (15)$$

With the additional assumption that the transition PDF for new samples only depend on the most recent parameters and observations, $q(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{0:t-1}, \mathbf{y}_{1:t}) \equiv q(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$, Equation 15 is rewritten as

$$w_t \propto \frac{p(\mathbf{y}_t | \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_t) \cdot p(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})}{q(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)} \cdot w_{t-1}. \quad (16)$$

The choice for the proposal distribution $q(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$ can be, among other, the *prior sampling* or the *optimal sampling* [37]. The prior sampling can lead to inefficient exploration because it does not take into account the current observation \mathbf{y}_t . Contrarily, in the optimal sampling, the current observation is included for the proposal, yet it is commonly difficult to sample from this proposal distribution because of its analytical intractability [29]. The limitation can be addressed using *auxiliary sampling* method [37] that resamples the particles using their marginal likelihood. The idea behind this sampling technique consists in preselecting the particles $\boldsymbol{\theta}_{t-1}$, that is, *surviving particles* that are likely to evolve

into highly plausible particles $\boldsymbol{\theta}_t$ by considering the current observation. The proposal distribution is defined as

$$q(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}^k, \mathbf{y}_t) = p(\mathbf{y}_t | \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{t-1}^k) \cdot p(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}^k) \cdot w_{t-1}^k, \quad (17)$$

where $k \in \mathcal{K} = \{1, 2, \dots, K\}$ is the auxiliary index of the particle at time $t - 1$. Therefore, the importance weight w_t^k in Equation 16 becomes

$$w_t^k \propto \frac{p(\mathbf{y}_t | \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_t^k)}{p(\mathbf{y}_t | \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{t-1}^k)}. \quad (18)$$

Because the auxiliary sampling method only prioritizes the surviving particles, it is prone to the diversity loss in the particles over time. This issue can be tackled by adding artificial noise to the particles in order to increase the exploration capacity [23]. Assuming that there is a set of particles at the time t , $\boldsymbol{\theta}_t^{\mathcal{K}} = \{\boldsymbol{\theta}_t^1, \boldsymbol{\theta}_t^2, \dots, \boldsymbol{\theta}_t^K\}$, the transition model for the particles is defined as

$$\boldsymbol{\theta}_t^k = \boldsymbol{\theta}_{t-1}^k + \mathbf{u}_t, \quad (19)$$

where \mathbf{u}_t is assumed to be a multivariate Gaussian distribution with mean zero and covariance matrix \mathbf{D}_t . Assuming that the artificial noise associated with each model parameter in the particle $\boldsymbol{\theta}_t^k$ is independent from each other, \mathbf{D}_t becomes diagonal matrix following

$$\mathbf{D}_t = \text{diag}(\underbrace{[\sigma_{u,1}^2 \ \sigma_{u,2}^2 \ \dots \ \sigma_{u,p}^2]^\top}_{\sigma_{u,t}^2}), \quad (20)$$

where P is defined as the number of unknown model parameters in the particle $\boldsymbol{\theta}_t^k$ and $\sigma_{u,p}$ corresponds to the standard deviation of the artificial noise for the p^{th} model parameter of $\boldsymbol{\theta}_t^k$. $\sigma_{u,t}$ are unknown hyperparameters to be estimated from data. Note that the introduction of the artificial noise leads to a loss of information [23]. The main steps of the RBPF are summarized in Algorithm 1.

Algorithm 1: Rao-Blackwellized Particle Filter (RBPF)

```

1 Given  $\boldsymbol{\theta}_0^{\mathcal{K}} \sim p(\boldsymbol{\theta}_0)$ ,  $\mathbf{D}_t$ ,  $\mathcal{K} = \{1, 2, \dots, K\}$ ,  $\mathbf{w}_0 = \frac{1}{K}$ ,  $\eta$ ;
2 for  $t = 1 : T$  do
3   for  $k = 1 : K$  do
4      $(\sim, \sim, \tilde{\mathcal{L}}_t^k, \sim) = \text{SKF}(\boldsymbol{\mu}_{t-1|t-1}^k, \boldsymbol{\Sigma}_{t-1|t-1}^k, \mathbf{y}_t, \mathbf{A}_{t-1}^k, \mathbf{C}_{t-1}^k, \mathbf{Q}_{t-1}^k, \mathbf{R}_{t-1}^k, \mathbf{Z}_{t-1}^k, \pi_{t-1|t-1}^k);$ 
5     Sample indices  $\mathcal{P}$  from  $\mathcal{K}$  with probability proportional to  $\mathcal{L}_t \odot \mathbf{w}_{t-1}$ ;
6      $\boldsymbol{\theta}_{t-1}^{\mathcal{K}} = \boldsymbol{\theta}_{t-1}^{\mathcal{P}}, \boldsymbol{\mu}_{t-1|t-1}^{\mathcal{K}} = \boldsymbol{\mu}_{t-1|t-1}^{\mathcal{P}}, \boldsymbol{\Sigma}_{t-1|t-1}^{\mathcal{K}} = \boldsymbol{\Sigma}_{t-1|t-1}^{\mathcal{P}}, \pi_{t-1|t-1}^{\mathcal{K}} = \pi_{t-1|t-1}^{\mathcal{P}}, \tilde{\mathcal{L}}_t^{\mathcal{K}} = \tilde{\mathcal{L}}_t^{\mathcal{P}};$ 
7     for  $k = 1 : K$  do
8        $\boldsymbol{\theta}_t^k = \boldsymbol{\theta}_{t-1}^k + \mathbf{u}_t, \mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_t);$ 
9        $(\boldsymbol{\mu}_{t|t}^k, \boldsymbol{\Sigma}_{t|t}^k, \mathcal{L}_t^k, \pi_{t|t}^k) = \text{SKF}(\boldsymbol{\mu}_{t-1|t-1}^k, \boldsymbol{\Sigma}_{t-1|t-1}^k, \mathbf{y}_t, \mathbf{A}_t^k, \mathbf{C}_t^k, \mathbf{Q}_t^k, \mathbf{R}_t^k, \mathbf{Z}_t^k, \pi_{t-1|t-1}^k);$ 
10      Compute weight  $w_t^k = \frac{\mathcal{L}_t^k}{\tilde{\mathcal{L}}_t^{\mathcal{K}}};$ 
11      Normalize weights  $\mathbf{w}_t = \frac{\mathbf{w}_t}{\sum_{k=1}^K w_t^k};$ 

```

3.2 Framework Architecture

This section presents the framework architecture for BDLM’s online learning procedure. Figure 1 illustrates the entire workflow. The framework architecture is separated into three

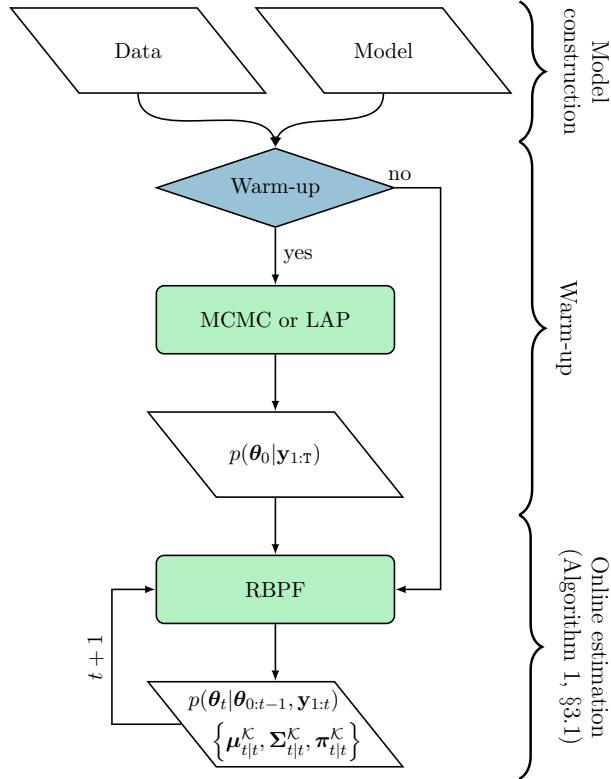


Figure 1: Illustration of the general framework of the online learning for the Bayesian Dynamic Linear Models. RBPF: Rao-Blackwellized Particle Filter; MCMC: Markov Chain Monte Carlo; LAP: Laplace Approximation.

main steps: *model construction*, *warm-up*, and *online estimation*. The model construction consists in pre-defining a vector of hidden state variables included in the model for interpreting the data. Examples of model construction are illustrated in several case-studies [30, 33]. The warm-up is employed for approximating the initial distribution for each model parameter. For this purpose, it can employ either the *Markov Chain Monte Carlo* (MCMC) or *Laplace Approximation* (LAP). The details for the application of these methods to BDLMs are described by Nguyen et al. [30]. This step ensures that the algorithm does not waste particles at places where the model parameter values are unlikely. Note that the warm-up step is operated in a batch learning procedure with a small training period. The online estimation is performed by the RBPF, as presented in Algorithm 1 (§3.1). Because each particle, $k \in \mathcal{K} = \{1, 2, 3, \dots, K\}$, represents a realization of the posterior PDF, there are K Gaussian PDF of the hidden state variables, as illustrated in Figure 2. Theoretically, the posterior predictive PDF for hidden state variables is obtained by integrating over the

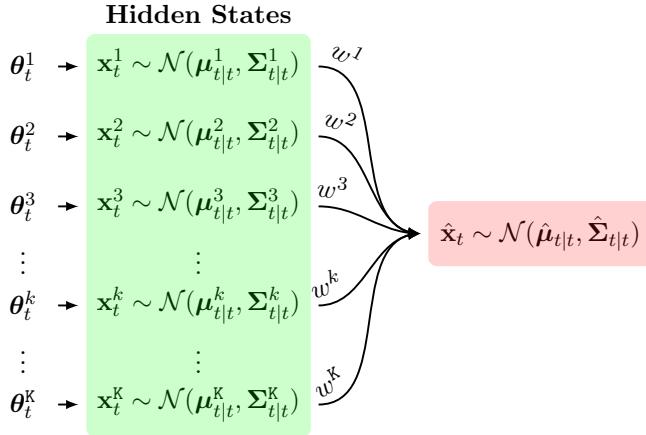


Figure 2: Illustration of the Gaussian mixture for the hidden state variables.

vector of the model parameters, $\boldsymbol{\theta}_t$,

$$p(\mathbf{x}_t | \mathbf{y}_{1:t}) = \int p(\mathbf{x}_t | \boldsymbol{\theta}_t, \mathbf{y}_{1:t}) \cdot p(\boldsymbol{\theta}_t | \mathbf{y}_{1:t}) d\boldsymbol{\theta}_t. \quad (21)$$

The vector of hidden state variables is a multivariate Gaussian distribution, as shown in Equation 3. The Gaussian mixture reduction [38] can be used to approximate the posterior predictive PDF of the hidden state variables using a single multivariate Gaussian PDF whose mean and covariance matrix are defined as

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{t|t} &= \sum_{k=1}^K \boldsymbol{\mu}_{t|t}^k \cdot w^k \\ \hat{\Sigma}_{t|t} &= \sum_{k=1}^K \boldsymbol{\Sigma}_{t|t}^k \cdot w^k + \sum_{k=1}^K (\boldsymbol{\mu}_{t|t}^k - \hat{\boldsymbol{\mu}}_{t|t})(\boldsymbol{\mu}_{t|t}^k - \hat{\boldsymbol{\mu}}_{t|t})^\top \cdot w^k, \end{aligned} \quad (22)$$

where w_t^k is the normalized importance weight of the particle $\boldsymbol{\theta}_t^k$. The online-estimation step is recursively repeated as each new data point arrives. Note that the warm-up step is optional because in some cases, there are no data available for gaining the prior knowledge about the model parameters. Hence, the online estimation step can be performed as a new observation is available.

In the context of anomaly detection, the model parameters in the BDLMs are categorized into *stationary* and *non-stationary* model parameters. The stationary model parameters denoted as $\boldsymbol{\theta}_t^s$, are constant over time. For these model parameters, the introduction of artificial noise, as presented in Section 3.1, can cause a high variability in the hidden state estimation. Therefore, the standard deviations of the artificial noise, $\sigma_{\mathbf{u},t}^s$, need to vanish overtime, so that

$$\sigma_{\mathbf{u},t}^s = \frac{1}{\alpha} \cdot \sigma_{\mathbf{u},t-1}^s, \quad (23)$$

where $\alpha > 1$ is a time-scaling factor. The non-stationary model parameters denoted as $\boldsymbol{\theta}_t^d$, are time-varying quantities that allow BDLMs to adapt to the changes of the underlying process in the data such as the occurrence of an abnormal event. A key challenge is that

the model parameters $\boldsymbol{\theta}_t^d$ struggle to adapt to the situation where the underlying process in data goes from one regime to another. In such a case, $\boldsymbol{\theta}_t^d$ tends to be stuck to the values of a single regime. This limitation leads to an increase in the variability in the hidden state estimation. Hence, it can jeopardize the timing as well as the accuracy of anomaly detection. To address this challenge, an initialization of these model parameters is triggered when the following conditions are satisfied

$$\begin{cases} \pi_{t-1|t-1}(\text{abnormal}) \geq \zeta \\ \pi_{t|t}(\text{abnormal}) < \zeta, \end{cases} \quad (24)$$

where $\pi_{t|t}(\text{abnormal})$ is the probability of the abnormal state at the time t and $\zeta \in (0, 1)$ is a probability threshold. In addition, the artificial noises of the stationary model parameters $\boldsymbol{\theta}_t^s$ need to be increased in order to provide a better exploitation. In this paper, these artificial noises are set to its initial values, $\sigma_{\mathbf{u},0}^s$, when the conditions in Equation 24 are met.

In the BDLM online learning framework, the standard deviations for the artificial noise, $\sigma_{\mathbf{u},0}$, are initialized based on the variance of the initial distribution for each model parameter. Each model-parameter group is defined following

$$\begin{aligned} \sigma_{\mathbf{u},0}^s &= \gamma^s \sqrt{\text{Var}(\boldsymbol{\theta}_0^s)} \\ \sigma_{\mathbf{u},0}^d &= \gamma^d \sqrt{\text{Var}(\boldsymbol{\theta}_0^d)}, \end{aligned} \quad (25)$$

where γ^s , γ^d are the scaling- factor vectors for the stationary and non-stationary model parameters, respectively. In practice, the chosen values of γ^d are usually greater than γ^s and are assumed to be constant overtime. By increasing the capacity of exploration for the non-stationary model parameters, it allows the model to rapidly adapt to the changes in the observations.

4 Case Studies

The section presents the application of the real-time anomaly detection methodology to two case-studies. The first case-study employs the raw displacement data recorded on a Dam in Canada. The second case-study is conducted on the same dataset as the first case-study, except that two artificial anomalies are introduced to this dataset for the validation purpose.

4.1 Data Description

For the first case-study, the new approach is applied to the X-displacement data measured using an inverted pendulum. The location of this pendulum is presented in Figure 3. The entire dataset of the displacement is illustrated in Figure 4. The displacement shows a non-harmonic periodic pattern as well as a descending trend overtime. This periodic pattern reaches its maxima during winter and its minima during summer. The displacement data are recorded from 2002 to 2016 with a total of 8634 data points with a non-uniform time-step length. Figure 5 shows the time-step length for the entire dataset. The most frequent

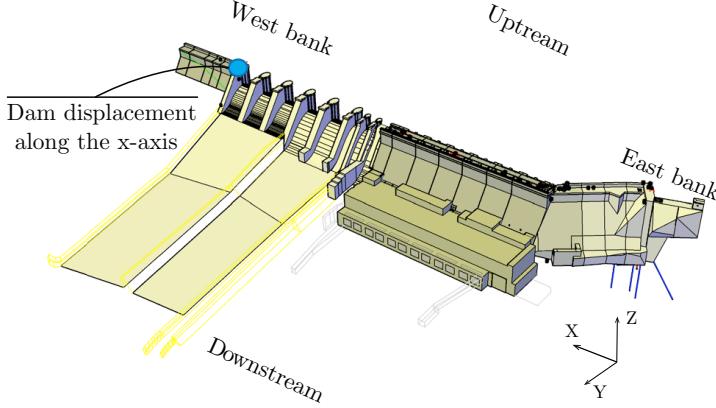


Figure 3: The location of the inverted pendulum.

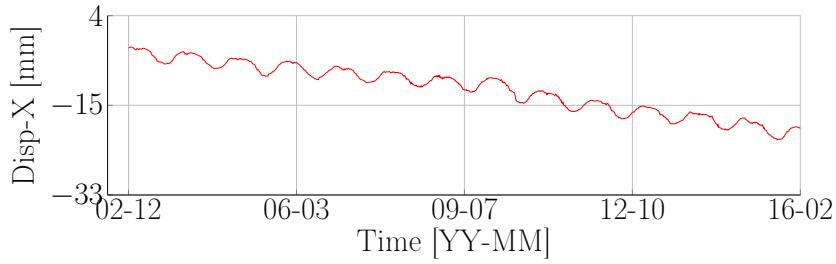


Figure 4: Illustration of the raw displacement data in X-direction.

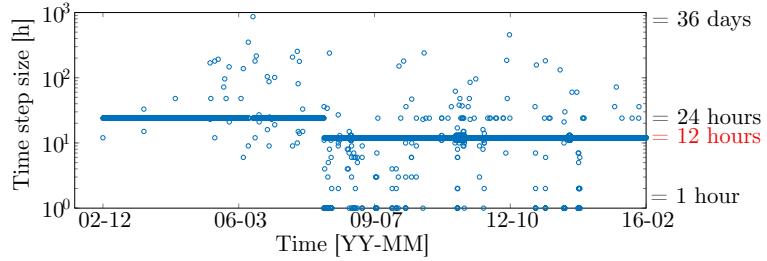


Figure 5: The time-step length for the X-displacement data.

time-step length is 12 hours. The observation error, σ_v , provided by the instrumentation engineers, is fixed to 0.3 mm.

The second case-study studies the same dataset as the first case-study except that two artificial anomalies are introduced to this dataset. The first anomaly is added from January 10 to February 6, 2009 with a slope of 15 % of the displacement baseline. The second one is introduced from July 12 to August 1, 2011 with the same slope as the previous one. The amount of data points for each anomaly-period (≈ 1 month) is 60 data points. Figure 6 shows the superposition of the original and artificial-anomaly dataset.

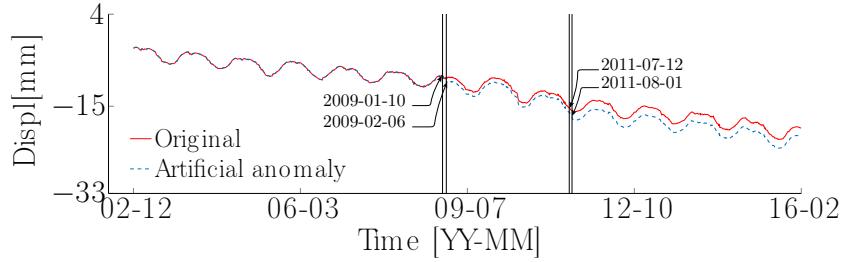


Figure 6: The superposition of the original and artificial-anomaly dataset.

4.2 Model Construction

The model of both case-studies consists in a vector of hidden state variables that includes a baseline (B), a local trend (LT), a local acceleration (LA), a kernel-regression (KR) component [31] with a period of 365.24 days, and an autoregressive (AR) component. The structural behavior over time is described by the baseline. The local trend is employed to model the rate of changes in the baseline. The local acceleration is used to model the rate of changes in the local trend. The kernel-regression component including 11 hidden state variables, describes the non-harmonic periodic pattern. The autoregressive component is used to capture the time-dependent model errors. The vector of hidden state variables is written as

$$\mathbf{x}_t = [x_t^B \ x_t^{LT} \ x_t^{LA} \ x_t^{KR} \ x_{t,0}^{KR} \ x_{t,1}^{KR} \ \dots \ x_{t,10}^{KR} \ x_t^{AR}]^\top. \quad (26)$$

Because the main interest here is to detect anomalies for the displacement data, two model classes representing the normal and abnormal states (see §2.3) are defined for this model. These two model classes use the same vector of hidden state variables presented in Equation 26, except that the local acceleration of the normal model-class is set to zero.

The unknown model parameters relating to the hidden state variables are defined as

$$\boldsymbol{\theta}_t = \left[\underbrace{\ell^{KR} \ \phi^{AR} \ \sigma_w^{AR}}_{\boldsymbol{\theta}_t^s} \ \underbrace{Z^{11} \ \sigma_w^{LA} \ Z^{22}}_{\boldsymbol{\theta}_t^d} \right]^\top, \quad (27)$$

where ℓ^{KR} is the kernel lengthscale, ϕ^{AR} is the autoregression coefficient, σ_w^{AR} is the autoregression standard deviation, Z^{ii} is the transition probability, and σ_w^{LA} is the local acceleration standard deviation. In this model, $\{\ell^{KR}, \phi^{AR}, \sigma_w^{AR}, Z^{11}\}$ and $\{\sigma_w^{LA}, Z^{22}\}$ are defined as the stationary and non-stationary model parameters, respectively. The kernel length-scale, standard deviation, are real number, \mathbb{R}^+ . The transition probability and autoregression coefficient are constrained to the interval $[0, 1]$. For an efficient learning procedure, the model parameters are transformed into the unbounded space. The standard deviations and kernel length-scale employs the natural logarithm as the transformation function. The sigmoid function is applied to the transition probabilities and autoregression coefficient.

In both case-studies, the Laplace Approximation (LAP) is employed for estimating the initial distribution for each model parameter with a training period of 1024 days (1004 data

points). The initial values for model parameter are defined using expert judgment and experience as well as prior data analysis

$$\boldsymbol{\theta}_0 = \begin{bmatrix} 0.5 & 0.95 & 0.095 & 0.9999 & 10^{-7} & 0.95 \\ \ell^{\text{KR}} & \phi^{\text{AR}} & \sigma_w^{\text{AR}} & Z^{11} & \sigma_w^{\text{LA}} & Z^{22} \end{bmatrix}^\top, \quad (28)$$

The complete model matrices are detailed in Appendix A. The hyperparameters, as presented in Section 3.2, are tuned as follows $\alpha = 1.01$, $\zeta = 0.5$, $\gamma^s = 0.01$, $\gamma^d = 0.1$. These hyperparameter are obtained using the empirical study where the different values of the hyperparameter are tested on multiple datasets. The number of particles for both case-studies is 60 000. The computing task is accelerated with *Graphics Processing Unit* (GPU).

4.3 Results

This section provides the results of anomaly detection and the estimation of hidden state variables as well as model parameters in for both case-studies.

4.3.1 Case-study #1

The new approach identifies that there was an anomaly occurring on July 9, 2010. This anomaly was due to the refection work on the dam in early July. Figure 7 presents the probabilities of the abnormal state for the displacement data over time. The solid line

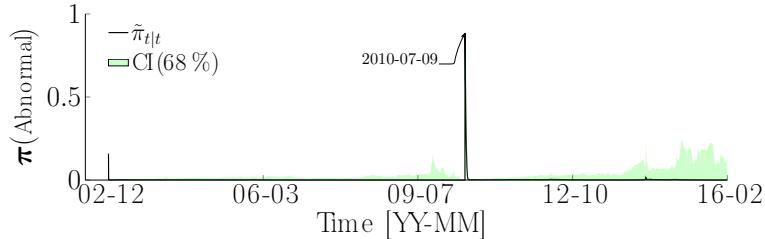


Figure 7: Probability of abnormal state for the displacement data in the case study #1.

presents the median values of the abnormal state probability, $\tilde{\pi}_{t|t}$. The shaded region presents its CI(68 %). The displacement behavior returns to the normal state once the work is completed. The estimation of the hidden state variables is presented in Figure 8. The mean values, $\hat{\mu}_{t|t}$, and its uncertainty bound, $\hat{\mu}_{t|t} \pm \hat{\sigma}_{t|t}$, are presented by the solid line and shaded region, respectively. The anomaly causes a abrupt change in the baseline (\mathbf{x}^B), local trend (\mathbf{x}^{LT}), and local acceleration (\mathbf{x}^{LA}), as illustrated in Figure 8a, b, and c, respectively. The lack of information at an early stage causes a large variability in the hidden state variables. This variability vanishes over time as more data are collected. Figure 8d shows that the model succeeds in separating the structural behavior from the periodic external effect. The autoregressive component presented in Figure 8e, shows a stationary process, as expected.

Figure 9 presents the Kernel smoothing function estimate [5] of the PDFs for each model parameter at 3 am on June 30, 2014. Figure 10 presents the evolution of the model

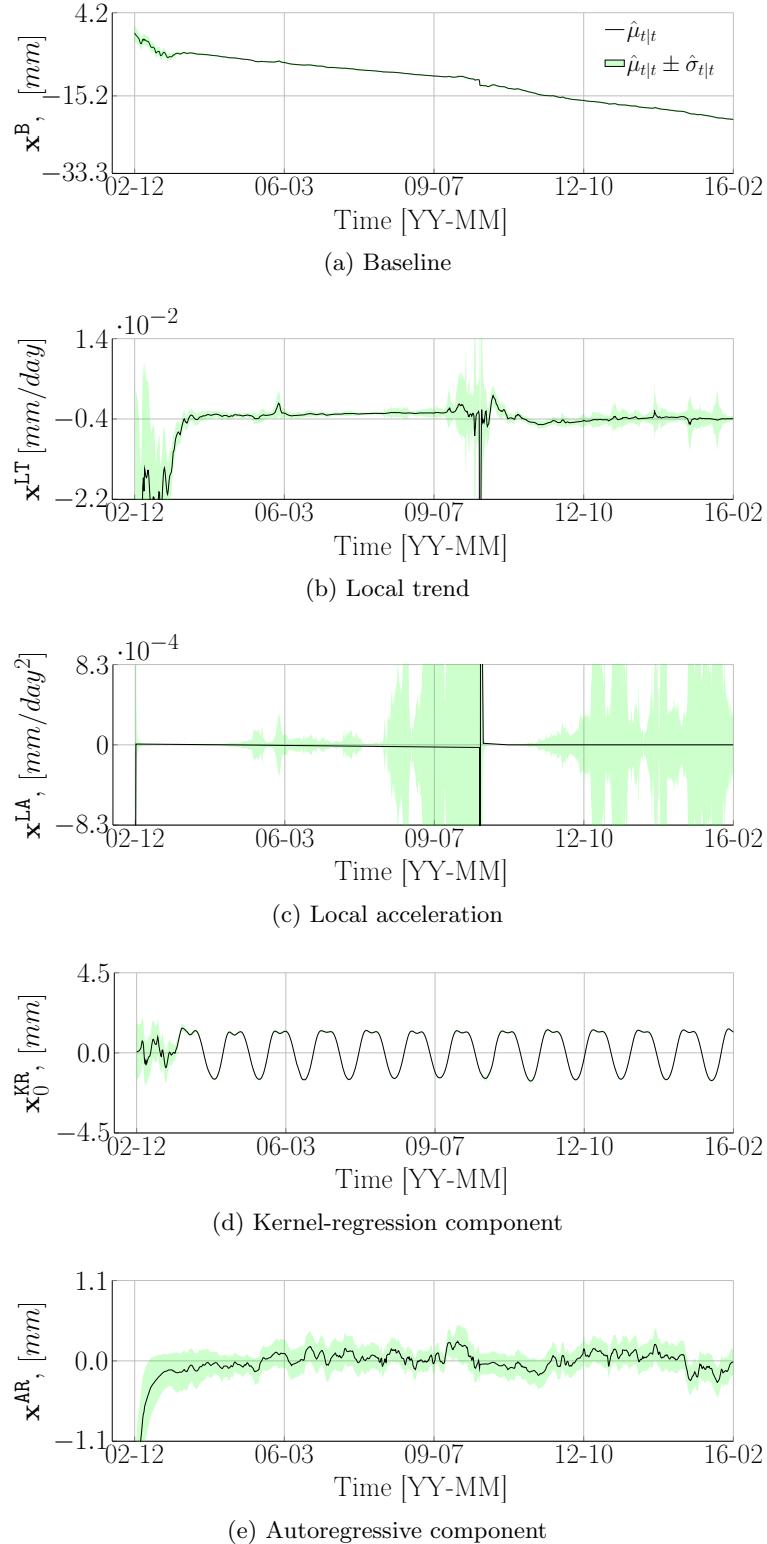


Figure 8: Illustration of the estimation of the hidden state variables for the case study #1.

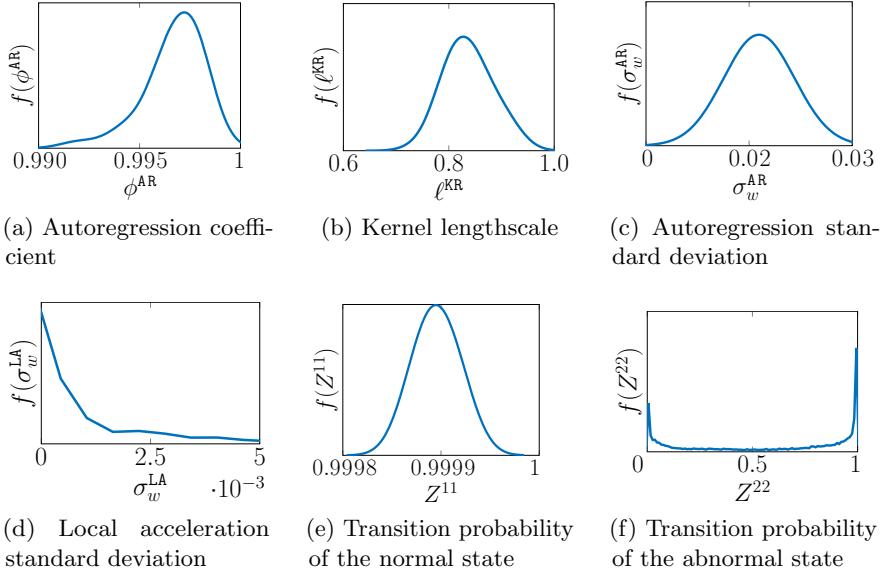


Figure 9: Kernel smoothing function estimate of the posterior PDFs for each model parameter at 3 am on June 30, 2014.

parameters over time. The median values, $\tilde{\theta}_t$, and its 68 % confidence interval, are presented by the solid line and shaded region, respectively. The large uncertainty in the stationary model parameters during the initial period is due to the imperfect initial conditions, as illustrated in Figure 10 a, b, c, and e. It then reduces over time with an increase in data points. When the anomaly took place, there is an abrupt change across all model parameters. However, the most significant change is found in the local acceleration standard deviation, σ_w^{LA} , relating to the abnormal state. When the abnormal events are absent, a large variability can be observed in the non-stationary model parameters such as Z^{22} and σ_w^{LA} , as presented in Figure 10 d. This behavior can be explained by a heavy tailed distribution of σ_w^{LA} and the presence of the bimodal distribution in Z^{11} , as illustrated in Figure 9 d and f.

4.3.2 Case-study #2

The anomaly occurring on July 9, 2010 as shown in case-study #1, is found in this case-study. In addition to this anomaly, two other anomalies are identified on February 6, 2009 and August 6, 2011. These two anomalies correspond to the periods where the artificial anomalies were introduced to the original dataset (see §4.1). Figure 11 presents the probability of abnormal state for the displacement dataset over time. The median values of the abnormal state probability, $\tilde{\pi}_{t|t}$, and their 68 % confidence interval, are presented by the solid line and shaded region, respectively. The probability of abnormal state of the artificial anomaly is presented by the dashed line. Because gradually adding a small slope to the original data takes time to create a significant change in their underlying process, the timing of anomaly detection provided by the model, does not match with the starting dates of the artificial anomaly. Figure 12 presents the hidden state variables for the entire dataset. The solid

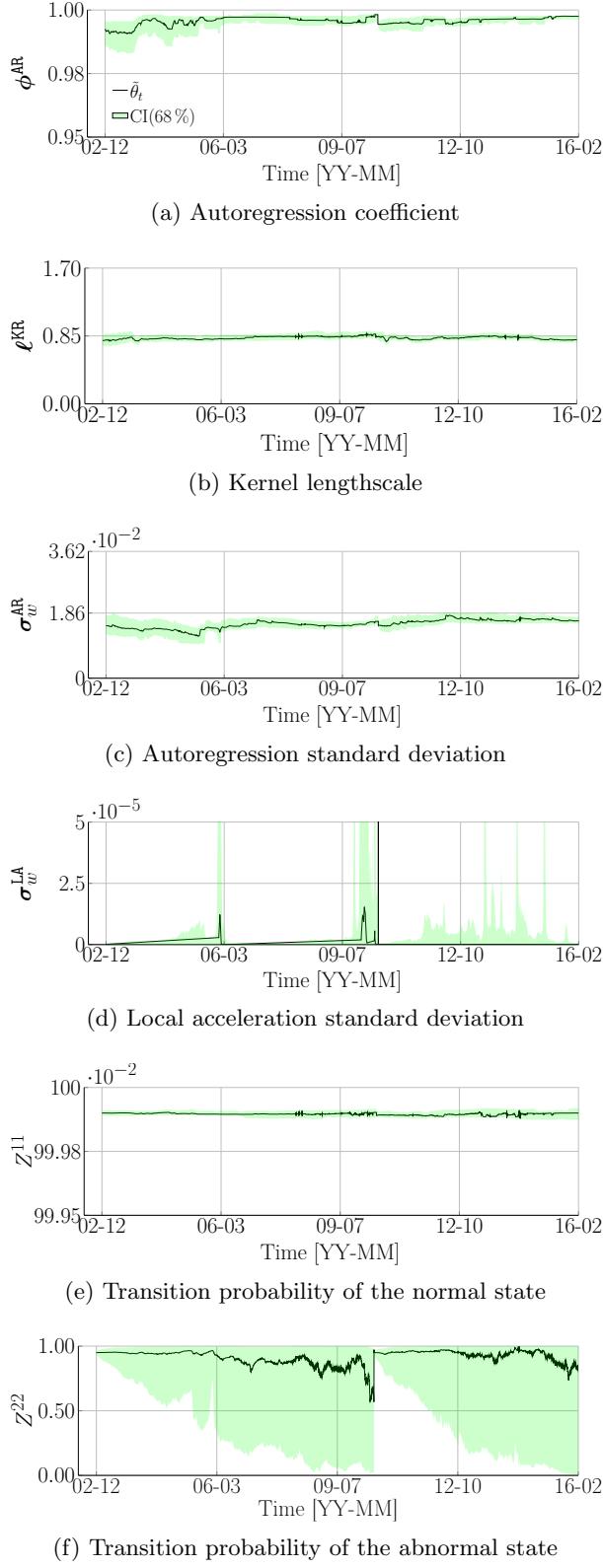


Figure 10: Illustration of the estimation of model parameters using Rao-Blackwellized particle filter for case-study #1.

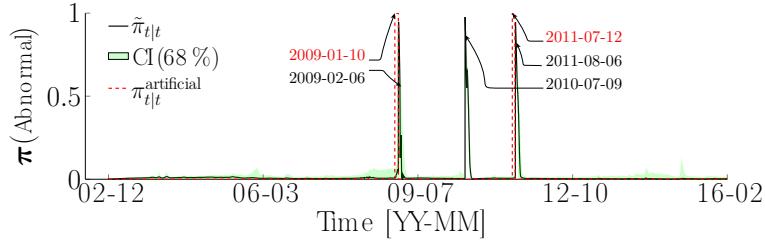


Figure 11: Probability of abnormal state for the displacement data in the case study #2.

line and shaded region represents the mean values, $\hat{\mu}_{t|t}$, and uncertainty bound, $\hat{\mu}_{t|t} \pm \hat{\sigma}_{t|t}$, accordingly. As case-study #1, the sudden changes are present in the baseline, local trend, and local acceleration when the anomalies occurred. The large variability across all hidden state variables during the initial period decreases over time as more data are available. Despite the presence of multiple anomalies, the model is capable of separating the structural behavior from the external effect, as illustrated in Figure 12d. Also, the autoregressive component follows a stationary process, as presented in Figure 12e.

The estimation of model parameters for the case-study #2 is presented in Figure 13. The median values of the model parameters, $\tilde{\theta}_t$, and its 68 % confidence interval, are presented by the solid line and shaded region, respectively. The same behavior as the case-study #1, has been found in the case-study #2. The large variability in the stationary model parameters is observed at an early stage. It then vanishes as more data are collected. A sudden change is identified across the model parameters. Three major changes corresponding to three abnormal events are found in the local acceleration standard deviation. As the case-study #1, a large uncertainty during the normal state is also identified in the transition probability of the abnormal state and local acceleration standard deviation.

5 Discussion

The potential of the new methodology for detecting anomalies in real-time is illustrated on both case-studies. The new approach is capable of detecting the anomaly caused by the refection work and also the anomalies being artificially introduced to the original dataset. In addition to the real-time anomaly detection, the proposed method allows estimating the hidden state variables as well as model parameters as the data are collected. Because the approach is accelerated with GPU computation, the estimation task of each time step for both case-studies was completed within approximately 1 second. This computational time is negligible in comparison with the data collection frequency in which the most frequent time-step is 12 hours, as shown in Figure 5.

In order to make an efficient performance, the initial distribution of model parameters for both case-studies are estimated using Laplace approximation with 1004 data points (1024 days). Also, the period of the external effect is a known quantity. Therefore, a small dataset should be available for carrying on the prior data analysis. As presented in Section 3.2, the hyperparameters need to be tuned before performing the anomaly-detection task. The generic hyperparameters can be obtained using the empirical study. The observation

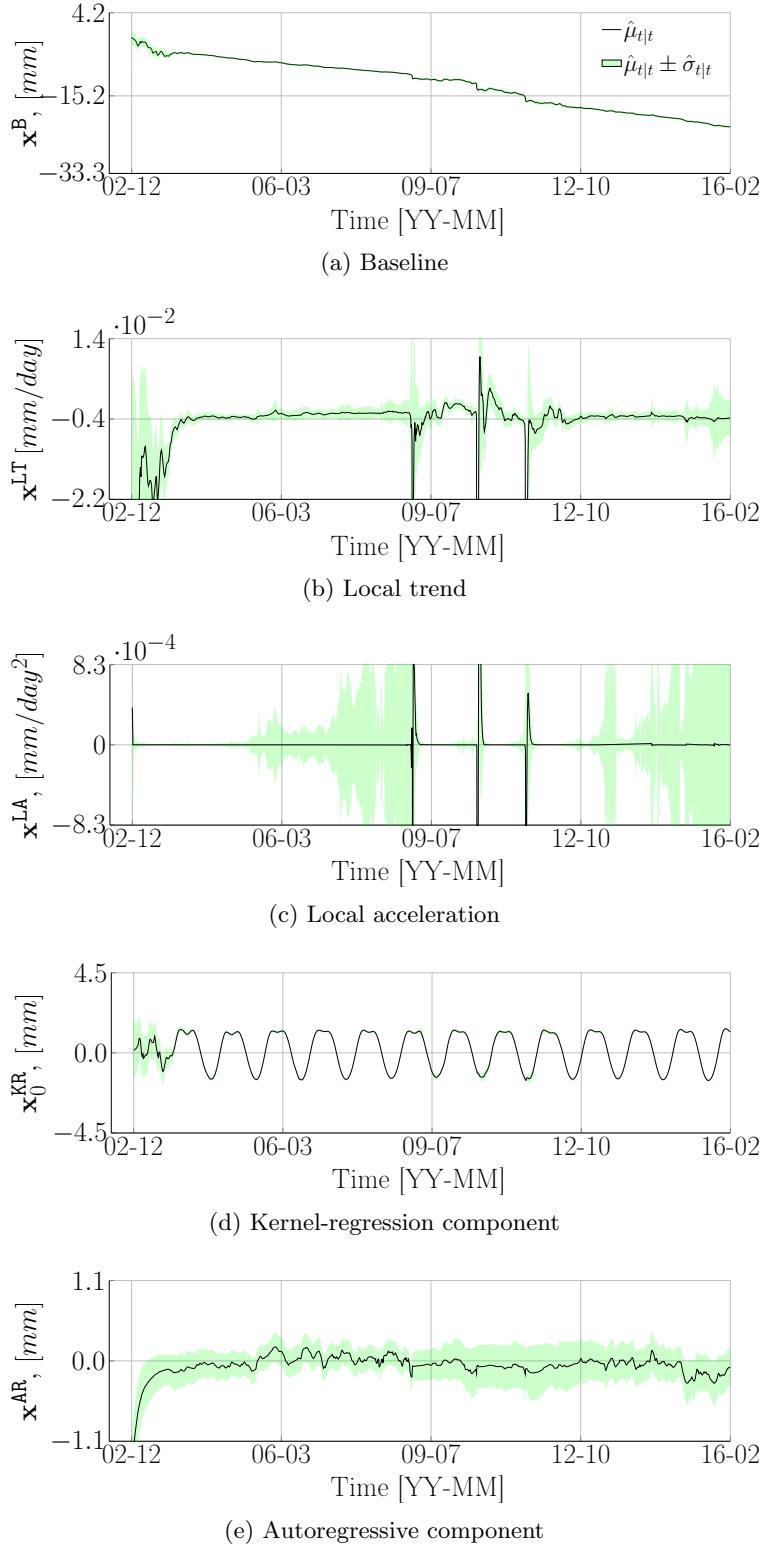


Figure 12: Illustration of the estimation of the hidden state variables for the case study #2.

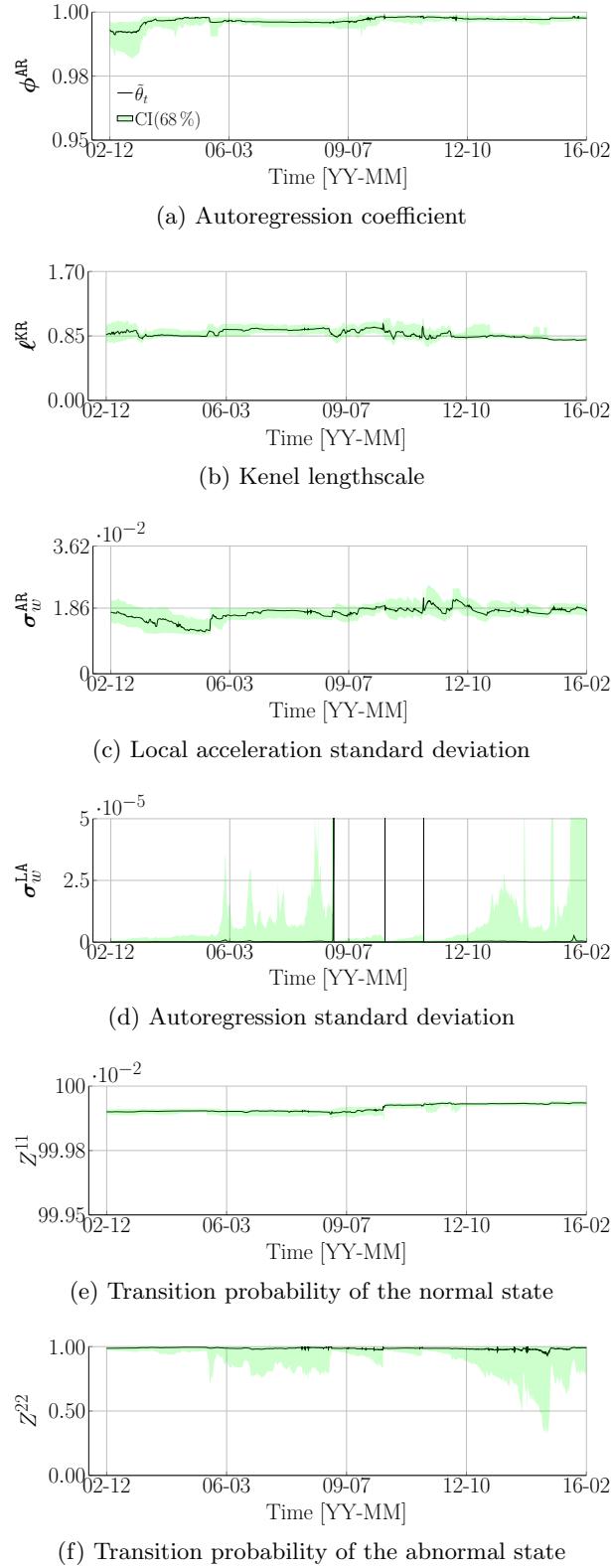


Figure 13: Illustration of the estimation of model parameters using Rao-Blackwellized particle filter for case-study #2.

error standard deviation, σ_v , in both case-studies is assumed to be constant over time. In practices, this hypothesis might no longer be valid because of the presence of either the sensor drift or the imperfect installation conditions. Future work should investigate the possibility of taking into account such these phenomena in the current framework in order to reduce the false alarm.

6 Conclusion

The paper proposes a new approach by combining the existing Bayesian dynamic linear models with the Rao-Blackwellized particle filter for the real-time anomaly detection. This approach allows estimating the hidden state variables and the model parameters as a new data point becomes available. The potential of the new method is illustrated on the displacement data recorded on a Dam in Canada. The method succeeds in detecting the anomaly due to the refection work on the dam as well as the artificial anomalies being introduced to the original dataset. The new method opens the way for monitoring the structure's health and conditions in real-time.

Acknowledgements

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 Benjamin Miquel and Vincent Roy from Hydro Québec for their help in the project.

References

- [1] R. Prescott Adams and D. JC. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.
- [2] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In *Proceedings of the 29th international conference on Very large data bases-Volume 29*, pages 81–92. VLDB Endowment, 2003.
- [3] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134–147, 2017.
- [4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.
- [5] A. W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford, New York, 1997.
- [6] F. Cadini, C. Sbarufatti, M. Corbetta, and M. Giglio. A particle filter-based model selection algorithm for fatigue damage identification on aeronautical structures. *Structural Control and Health Monitoring*, 24(11):e2002, 2017.

- [7] F.N. Catbas, T. Kijewski-Correa, and A.E. Aktan. *Structural Identification of Constructed Facilities. Approaches, Methods and Technologies for Effective Practice of St-Id.* American Society of Civil Engineers (ASCE), Reston, VA, 2013.
- [8] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 143–152. IEEE, 2012.
- [9] E. N. Chatzi and A. W. Smyth. Particle filter scheme with mutation for the estimation of time-invariant parameters in structural health monitoring applications. *Structural Control and Health Monitoring*, 20(7):1081–1095, 2013.
- [10] Y. Cui, S. Ahmad, and J. Hawkins. Continuous online sequence learning with an unsupervised neural network model. *Neural computation*, 28(11):2474–2504, 2016.
- [11] A. Doucet, N. De Freitas, K. P. Murphy, and S. Russell. Rao-blackwellised particle filtering for dynamic bayesian networks. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 176–183. Morgan Kaufmann Publishers Inc., 2000.
- [12] J. Durbin and S. J. Koopman. *Time series analysis by state space methods*, volume 38. Oxford University Press, 2012.
- [13] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. CRC Press, 3 edition, 2014.
- [14] J.-A. Goulet. Bayesian dynamic linear models for structural health monitoring. *Structural Control and Health Monitoring*, 24:e2035–n/a, 2017. e2035 stc.2035.
- [15] G. Grisetti, C. Stachniss, and W. Burgard. Improving grid-based slam with rao-blackwellized particle filters by adaptive proposals and selective resampling. In *ICRA*, pages 2432–2437. ”, 2005.
- [16] G. Grisetti, G. D. Tipaldi, C. Stachniss, W. Burgard, and D. Nardi. Fast and accurate slam with rao–blackwellized particle filters. *Robotics and Autonomous Systems*, 55(1):30–38, 2007.
- [17] S. Guha, N. Mishra, G. Roy, and O. Schrijvers. Robust random cut forest based anomaly detection on streams. In *International conference on machine learning*, pages 2712–2721. ”, 2016.
- [18] D. J. Hill, B. S. Minsker, and E. Amir. Real-time bayesian anomaly detection for environmental sensor data. In *Proceedings of the Congress-International Association for Hydraulic Research*, volume 32, page 503. Citeseer, 2007.
- [19] M. Hoshiya and E. Saito. Structural identification by extended kalman filter. *Journal of Engineering Mechanics*, 110(12):1757–1770, 1984.

- [20] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.
- [21] D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pages 180–191. VLDB Endowment, 2004.
- [22] N. Laptev, J. Yosinski, L. E. Li, and S. Smyl. Time-series extreme event forecasting with neural networks at uber. In *International Conference on Machine Learning*, number 34, pages 1–5. ”, 2017.
- [23] J. Liu and M. West. Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo methods in practice*, pages 197–223. Springer, 2001.
- [24] J.P. Lynch and K.J. Loh. A summary review of wireless sensors and sensor networks for structural health monitoring. *Shock and Vibration Digest*, 38(2):91–130, 2006.
- [25] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal. Long short term memory networks for anomaly detection in time series. In *Proceedings*, page 89. Presses universitaires de Louvain, 2015.
- [26] O. Maruyama and M. Hoshiya. System identification of an experimental model by extended kalman filter. *Structural Safety and Reliability: ICOSSAR'01*, page 2001, 2001.
- [27] K. P. Murphy. Switching kalman filters. Technical report, Citeseer, 1998.
- [28] K. P. Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, London, UK, 2012.
- [29] C. Nemeth, P. Fearnhead, and L. Mihaylova. Sequential monte carlo methods for state and parameter estimation in abruptly changing environments. *arXiv preprint arXiv:1510.02604*, 2015.
- [30] L. H. Nguyen, I. Gaudot, and J.-A. Goulet. Uncertainty quantification for model parameters and hidden state variables in bayesian dynamic linear models. *Structural Control and Health Monitoring*, 26(3):e2309, 2019.
- [31] L. H. Nguyen, I. Gaudot, S. Khazaeli, and J.-A. Goulet. A kernel-based method for modeling non-harmonic periodic phenomena in bayesian dynamic linear models. *Frontiers in Built Environment*, 5:8, 2019.
- [32] L. H. Nguyen and J.-A. Goulet. Anomaly detection with the switching kalman filter for structural health monitoring. *Structural Control and Health Monitoring*, pages e2136–n/a, 2018. e2136 stc.2136.
- [33] L.H. Nguyen and J-A. Goulet. Structural health monitoring with dependence on non-harmonic periodic hidden covariates. *Engineering Structures*, 166:187 – 194, 2018.

- [34] A. Olivier and A. W. Smyth. On the performance of online parameter estimation algorithms in systems with various identifiability properties. *Frontiers in Built Environment*, 3:14, 2017.
- [35] A. Olivier and A. W. Smyth. Particle filtering and marginalization for parameter identification in structural systems. *Structural Control and Health Monitoring*, 24(3):e1874, 2017.
- [36] M.A.F. Pimentel, D.A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215 – 249, 2014.
- [37] M. K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599, 1999.
- [38] A. R. Runnalls. Kullback-leibler approach to gaussian mixture reduction. *IEEE Transactions on Aerospace and Electronic Systems*, 43(3):989–999, 2007.
- [39] S. Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, New York, US, 2013.
- [40] S. Särkkä, A. Vehtari, and J. Lampinen. Rao-blackwellized particle filter for multiple target tracking. *Information Fusion*, 8(1):2–15, 2007.
- [41] T. J. Sejnowski and C. R. Rosenberg. Parallel networks that learn to pronounce english text. *Complex systems*, 1(1):145–168, 1987.
- [42] M. Szmit and A. Szmit. Usage of modified holt-winters method in the anomaly detection of network traffic: Case studies. *Journal of Computer Networks and Communications*, 2012, 2012.
- [43] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics. Springer New York, New York, US, 1999.

Appendix A Model Matrices

The *transition matrix* (\mathbf{A}_t), the *observation matrix* (\mathbf{C}_t), the *observation error covariance matrix* (\mathbf{R}_t), and the *model error covariance matrix* (\mathbf{Q}_t) for normal model class and abnormal model class are defined following

Normal model class

$$\begin{aligned}\mathbf{A}_t^1 &= \text{block diag} \left(\begin{bmatrix} 1 & \Delta t & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & \tilde{\mathbf{k}}_t^{\text{KR}} \\ \mathbf{0} & \mathbf{I}_{11} \end{bmatrix}, \phi^{\text{AR}} \right) \\ \mathbf{C}_t^1 &= [1 \ 0 \ 0 \ 0_1 \ 0_2 \dots 0_{11} \ 1] \\ \mathbf{R}_t^1 &= [(\sigma_v)^2] \\ \mathbf{Q}_t^{1(1)} &= \text{block diag} (\mathbf{0}_3, \mathbf{0}_{11}, (\sigma^{\text{AR}})^2) \\ \mathbf{Q}_t^{2(1)} &= \text{block diag} (\mathbf{0}_3, \mathbf{0}_{11}, (\sigma^{\text{AR}})^2)\end{aligned}$$

Abnormal model class

$$\begin{aligned}\mathbf{A}_t^2 &= \text{block diag} \left(\begin{bmatrix} 1 & \Delta t & \frac{\Delta t^2}{2} \\ 0 & 1 & \Delta t \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & \tilde{\mathbf{k}}_t^{\text{KR}} \\ \mathbf{0} & \mathbf{I}_{10} \end{bmatrix}, \phi^{\text{AR}} \right) \\ \mathbf{C}_t^2 &= [1 \ 0 \ 0 \ 0_1 \ 0_2 \dots 0_{11} \ 1] \\ \mathbf{R}_t^2 &= [(\sigma_v)^2] \\ \mathbf{Q}_t^{1(2)} &= \text{block diag} (\mathbf{0}_3, \mathbf{0}_{11}, (\sigma^{\text{AR}})^2) \\ \mathbf{Q}_t^{2(2)} &= \text{block diag} \left((\sigma^{\text{LA}})^2 \cdot \begin{bmatrix} \frac{\Delta t^2}{20} & \frac{\Delta t^4}{8} & \frac{\Delta t^3}{6} \\ \frac{\Delta t^4}{8} & \frac{\Delta t^3}{3} & \frac{\Delta t^2}{2} \\ \frac{\Delta t^3}{6} & \frac{\Delta t^2}{2} & \Delta t \end{bmatrix}, \mathbf{0}_{11}, (\sigma^{\text{AR}})^2 \right),\end{aligned}$$

where $\tilde{\mathbf{k}}_t^{\text{KR}} = [\tilde{k}_{t,1}^{\text{KR}} \ \tilde{k}_{t,2}^{\text{KR}} \dots \tilde{k}_{t,11}^{\text{KR}}]$ is the normalized kernel values, Δt is the time step at the time t , \mathbf{I}_{10} is the 10×10 identity matrix, and $\mathbf{0}_n$ is the $n \times n$ zero matrix.