Bayesian Online State and Parameter Estimation for Streaming Data

Rui Vieira

Thesis submitted for the degree of
Doctor of Philosophy

School of Mathematics & Statistics
Newcastle University
Newcastle upon Tyne
United Kingdom

September 2018

Abstract

With the advent of Big Data and the Internet of Things, data streams are ubiquitous, increasing the demand for real-time inference on sequential data at low computational cost. Inference for streaming time-series is tightly coupled with the problem of Bayesian online state and parameter inference. In this thesis we will focus mainly on Dynamic Generalised Linear Models, the class of models often chosen to model continuous and discrete time-series data. We will look at methods which solve the problem of estimating jointly states and parameters, both in online and offline scenarios.

For the online scenario, when the parameters are known, we will look at the Kalman Filter and Sequential Monte Carlo methods (SMC) which provide estimations for the hidden latent states. We will then consider SMC extensions allowing for online joint state and parameter estimation.

Offline methods, by definition, do not allow real-time estimation but typically provide superior results at higher time and computational costs. In this thesis we propose and evaluate a fully online, approximated version of a sequential, but not-online method (SMC^2). This method approximates the true posterior, performing estimation over a sliding window of the most recent observations and so bounding the computational cost and operating in an online fashion, providing an acceptable approximation and, by employing particle rejuvenation through an MCMC move, delaying particle impoverishment problems.

This thesis analyses online methods when applied to different real world datasets showing that SMC sufficient statistics-based-methods delay known problems, such as particle impoverishment, especially when applied to long running time-series, while providing reasonable estimations when compared to exact methods, such as Particle Marginal Metropolis-Hastings. State and observation forecasts will also be analysed as a performance metric. By benchmarking against a "gold standard" (offline) method, we can better understand the performance of online methods in challenging real-world scenarios.

To Paula.

Acknowledgements

I would like to thank my supervisors, Prof. Darren Wilkinson and Prof. Paul Watson.

A special thank you to Prof. Darren Wilkinson for his inspiring views on the project, support and scientific expertise.

I would also like to express my gratitude to Red Hat Inc. for the financial support for this project. A special mention to Mark Little and Jonathan Halliday for supporting this project and their feedback.

Finally, I would like to thank Paula, my friends and my parents for their support and encouragement.

Contents

List of Figures

List of Tables

Part I

Dynamic Generalised Linear Models

Chapter 1

Introduction

With the modern ubiquity of large streaming datasets comes the requirement of robust real-time inference. A multitude of different data sources, such as Internet of Things (IoT) devices, server, network and sensor metrics, all exhibiting particular patterns and observation types, also increase the demand for flexible and computationally cheap solutions.

Some typical analyses performed on such streaming time-series are forecasting, anomaly detection and seasonal decomposition in order to perform statistically-based decisions typically under tight time constraints.

As standard offline methods, such as Markov chain Monte Carlo (MCMC), are normally not suitable when taking into account such constraints, here we analyse alternatives such as sequential Monte Carlo (SMC). Although SMC is well studied in the scientific literature and quite prevalent in academic research in the last decade, modern analytics platforms still resort to less powerful methods (such as moving averages).

When coupled with Dynamic Generalised Linear Models (DGLMs), a specific class of State Space Models, which allow to specify complex, linear and non-linear time-series patterns, SMC then enables performing real-time Bayesian estimations in a wide variety of streaming datasets.

Inference on streaming time-series is tightly coupled with the problem of Bayesian online state and parameter inference. In this thesis we will perform a review of some well established methods for SMC for DGLMs applied to distinct datasets. We will start by first introducing the DGLM, the class of state space models chosen for our data (Chapter 2). We will then introduce the fundamental theory of online state estimation (when the parameters are known) from a Bayesian perspective, usually called *Bayesian filtering* in Chapter 3. Direct applications of filtering methods to linear and non-linear DGLMs, namely the Kalman filter, Extended Kalman filter and conjugate filtering will be described in Chapters 4

and 5. Modern developments, such as the aforementioned SMC, are built in the theoretical concepts of *importance sampling* (IS) which will be detailed in Chapter 6, where different importance density alternatives will be presented. Chapter 7 will focus on the application of IS concepts to dynamic state-space models and the basic methods of SMC. In Part III we will look at extensions to SMC which allow for joint state and parameter estimation in DGLMs. Namely, in Chapter 8 we will look at state augmentation approaches (particularly the Liu and West filter) and in Chapters 9 and 10 methods which rely on the concept of sufficient statistics will be introduced.

In Part IV we will look at offline estimation methods. Kalman smoothing, in Chapter 11 will be analysed, since it is a fundamental concept which allows for the offline parameter estimation in Normal DLMs using the Expectation-Maximisation (EM) algorithm presented in Chapter 12. The method used as the "gold standard" to which all the methods presented in this thesis will be compared to is Particle Markov Chain Monte Carlo (PMMC), specifically Particle Marginal Metropolis-Hastings (PMMH) as presented in Chapter 13. Competing methods which provided sequential (but not online) estimation include the Iterated Batch Importance Sampling (IBIS) method (presented in Chapter 14) and SMC^2 (Chapter 15). *Ad-hoc* implementations which allow for an online (albeit approximate) joint state and parameter estimation for both methods will also be presented in these chapters.

Finally, in Part V we will analyse the performance of the online methods presented in terms of state and parameters estimation, computational cost, anomaly detection, short and long term state and observation forecast with four real-world datasets. We will also focus on topics which are directly relevant to the main application area which we approach, streaming time-series, such as the choice of resampler and the accumulation of Monte Carlo errors in long running time series.

Chapter 2

Dynamic Generalised Linear Models

2.1 State-space models

To model the data we chose the Dynamic Generalised Linear Model (DGLM), a specific instance of the more general class of State-Space Models (SSM), illustrated in Figure 2.1, where we have the relations

$$y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi}_t \sim f\left(y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi}_t\right) \tag{2.1}$$

$$\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi}_t \sim l\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi}_t\right). \tag{2.2}$$

Usually (2.1) is referred to as the *observational model* and (2.2) as the *system model*. We consider the discrete time case, $t \in \mathbb{N}$, and the state vector $\boldsymbol{\theta}_t \in \mathbb{R}^m$. $\boldsymbol{\Phi}_t$ is the set of parameters for this model and we will consider the special case where $\boldsymbol{\Phi}_t = \boldsymbol{\Phi}$ to be static (but not necessarily known) throughout. The sequence of state vectors Θ_t is a Markov

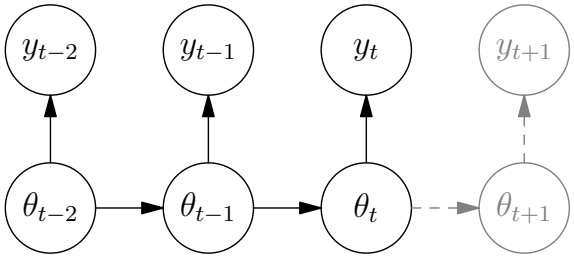


Figure 2.1: State-Space Model

Chain with transition density[1] l, such that

$$\Theta_t | \{\Theta_{t-1} = \boldsymbol{\theta}_{t-1}\} \sim l(\cdot|\boldsymbol{\theta}_{t-1})$$

and the sequence of observations $\mathcal{D}_t$ is the output of Θ_t such that $\mathcal{D}_t | \{\Theta_t = \boldsymbol{\theta}_t\} \sim f(\cdot|\boldsymbol{\theta}_t)$.

The second component, the *system model* (2.2), defined by the function $l : \mathbb{R}^m \mapsto \mathbb{R}^m$ can be non-linear although in DGLMs will be a linear function of the form

$$\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi} \sim p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi}) = \mathcal{N}(\boldsymbol{\theta}_t|\mathsf{G}_t\boldsymbol{\theta}_{t-1}, \mathsf{W}), \tag{2.3}$$

where the initial state is assumed to be distributed according to a normal, usually vague, prior

$$\boldsymbol{\theta}_0|\mathbf{m}_0, \mathsf{C}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0). \tag{2.4}$$

In (2.3), W (the state transition variance) and in (2.4), C_0 (the prior covariance), are both matrices of dimensions $m \times m$. In DGLMs the observational model, characterised by the density $f : \mathbb{R}^m \mapsto \mathbb{R}^n$, follows an exponential family distribution in the form of

$$y_t \sim p(y_t|\eta_t) = \exp\left\{\frac{z(y_t)\,\eta_t - b(\eta_t)}{a(\phi_t)} + c(y_t, \phi_t)\right\}. \tag{2.5}$$

We will consider throughout the case where $n = 1$, $f : \mathbb{R} \mapsto \mathbb{R}^+$ or $f : \mathbb{R} \mapsto [0, 1]$, respectively the continuous and discrete univariate case. In the literature (Dobson (2000)) η_t is usually called the *natural parameter* and ϕ_t the *dispersion parameter*. We consider $b(\cdot)$ to be convex and twice differentiable in η_t and that, as noted in West & Harrison (1997):

$$\mathrm{E}[z(y_t)|\eta_t] = \mu_t = \frac{db(\eta_t)}{d\eta_t} = b'(\eta_t) \tag{2.6}$$

$$\mathrm{Var}[z(y_t)|\eta_t] = \Sigma_t = \frac{a(\phi_t)\,d^2b(\eta_t)}{d\eta_t^2} = a(\phi_t)\,b''(\eta_t). \tag{2.7}$$

Other relations that will be used in later sections, include the *linear predictor* of the DGLM given by

$$\lambda_t = \mathsf{F}^T\boldsymbol{\theta}_t. \tag{2.8}$$

The relation between the mean presented in (2.6) and the linear predictor in (2.8) is given by the *response function* $h(\cdot)$

$$\mu_t = h(\lambda_t),$$

[1] In this thesis, l is used to refer to a measure or a density (with Lebesgue measure) depending on the context.

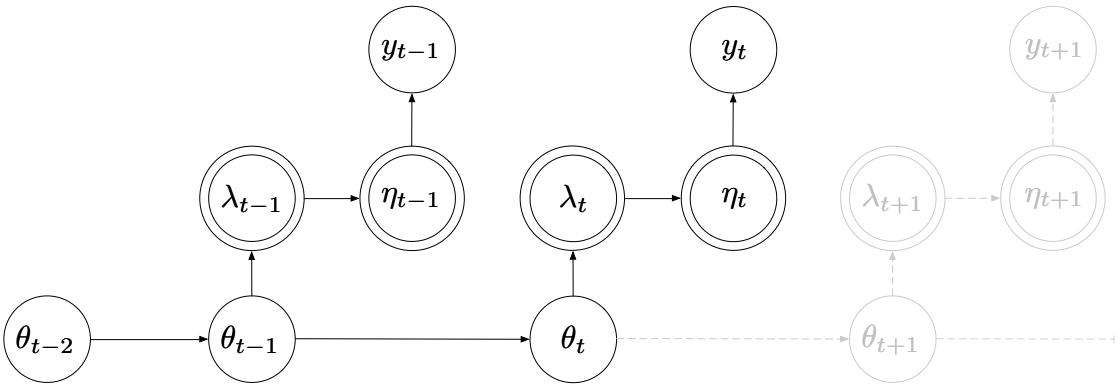


Figure 2.2: DGLM diagram. Double circles represent deterministic nodes.

and the inverse of the response function is usually called the *link function*

$$g(\mu_t) = h^{-1}(\mu_t) = \lambda_t. \tag{2.9}$$

For the remainder we will consider the *canonical link*, that is the specific instance where

$$\eta_t = \lambda_t. \tag{2.10}$$

The response and link function will be dependent on the specific distribution used for the observation model (2.1).

The factors F_t and G_t are respectively the *observation* and *system matrices* which allow us to specify the structure of our time series. These factors might represent a *locally constant model* (Section 2.2.1.1), where the states will represent an underlying mean, a *locally linear model* (Section 2.2.1.2), where the states represent a mean and a trend, or a seasonal model (Sections 2.2.2.1 and 2.2.2.2), where each component of the state will represent a cyclic component. As with the model's parameters, these matrices can vary in time but for the remainder of this thesis we will consider the case where they are static and known, that is $\mathsf{F}_t = \mathsf{F}$ and $\mathsf{G}_t = \mathsf{G}$, also referred as *time series models* in the literature (West & Harrison (1997)).

It is clear from the above definitions that this class of models possesses Markovian properties, that is, denoting the sequence of observations $y_{1:t-1}$ as $\mathcal{D}_{t-1}$ and omitting the dependence on the model parameters $\boldsymbol{\Phi}$:

$$p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_{t-1}) = p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}). \tag{2.11}$$

This represents the fact that future states will only depend on the present state (here represented as $\boldsymbol{\theta}_{t-1}$) and will be independent of past states, $\boldsymbol{\theta}_{0:t-2}$, and observations, $\mathcal{D}_{t-1}$

(as represented in Figure 2.2). Additionally, at any given time $t-1$, the state $\boldsymbol{\theta}_{t-1}$ will not depend on any future values, that is

$$p\left(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_{t:T}, y_{t:T}\right) = p\left(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_{t}\right). \tag{2.12}$$

Finally we can state that at any time t the observation y_t will only depend on the current state vector $\boldsymbol{\theta}_t$:

$$p\left(y_t|\boldsymbol{\theta}_{0:t}, \mathcal{D}_{t-1}\right) = p\left(y_t|\boldsymbol{\theta}_{t}\right). \tag{2.13}$$

DGLMs are a flexible and elegant tool for modelling streaming data, since they can represent both discrete and continuous data by appropriate selection of the observation model, as well as providing the means to express complex time-series behaviour by composing simpler ones. In the remainder of this thesis we will refer to a specific DGLM by classifying it according the observation model (Poisson, Binomial) as detailed in Section 2.3.

2.2 Model specification

2.2.1 n^{th}-order polynomial models

At this point, we introduce the concept of *Jordan blocks*. For any real or complex λ, a n-Jordan block is defined (West & Harrison (1997)) as

$$\mathbf{J}_n\left(\lambda\right) = \begin{bmatrix} \lambda & 1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & 0 & \dots & 0 \\ 0 & 0 & \lambda & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & 0 & \dots & \lambda \end{bmatrix}_{n\times n}.$$

That is, an $n \times n$ upper triangular diagonal matrix with diagonal elements of λ and super-diagonal elements 1.

In general terms an n^{th}-order polynomial model can be defined by the the following observation and system matrices

$$\begin{aligned} \mathsf{F} &= \underbrace{[1 \quad 0 \quad 0 \quad 0 \quad \cdots \quad 0]^T}_{n} \\ \mathsf{G} &= \mathbf{J}_n(1), \end{aligned}$$

with the state vector $\boldsymbol{\theta}_t$ having dimension n and the state transition variance, W, having dimension $n \times n$.

An n^{th}-order polynomial model will be referred to as $\mathcal{P}(n)$ and for the remainder of this thesis we will focus on two specific cases of these models, namely the *first-order* (*locally constant*, $\mathcal{P}(1)$) and *second-order* (*locally linear*, $\mathcal{P}(2)$) polynomial models.

2.2.1.1 Locally constant

The simplest polynomial model is the *locally constant* model, $\mathcal{P}(1)$, also referred in the literature as a *first-order polynomial*[2] model.

In this instance

$$\mathsf{F} = \mathsf{G} = \mathbf{J}_1(1) = [1].$$

The state transition in this case amount to a random walk expressed as

$$\begin{aligned} \theta_t &\sim \mathcal{N}\left(\theta_{t-1}, \tau^2\right) \\ &= \theta_{t-1} + \omega_t, \qquad \omega_t \sim \mathcal{N}\left(0, \tau^2\right). \end{aligned}$$

2.2.1.2 Locally linear

The *locally linear model*, referred to usually as the *second-order polynomial* model, $\mathcal{P}(2)$, can be defined as

$$\begin{aligned} \mathsf{F} &= [1 \quad 0]^T \\ \mathsf{G} &= \mathbf{J}_2(1) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}. \end{aligned}$$

In this case, the state vector $\boldsymbol{\theta}_t$ will consist of two elements, respectively $\boldsymbol{\theta}_t = \left(\boldsymbol{\theta}_t^{level} \quad \boldsymbol{\theta}_t^{trend}\right)$, a *level* and a *trend*.

[2] In this model, the underlying mean is locally constant. We will use the terminology of West *et al.* (1985) where this is considered a *first-order* polynomial model (*cf.* West *et al.* (1985), pp.32-34)

2.2.2 Seasonality

2.2.2.1 Full form

A basic representation for periodic behaviour (of a fixed period p), is to represent F and G as

$$\mathsf{F}= \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{\mathsf{p}} \qquad \mathsf{G} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}_{\mathsf{p}\times\mathsf{p}} .$$

2.2.2.2 Fourier components

An alternative form of expressing cycling behaviour in DLMs is to use a trigonometric representation, usually called *Fourier components* in the literature (West & Harrison, 1997).

Each harmonic i of the seasonal components is expressed as:

$$\mathsf{F}_i = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathsf{G}_i = \begin{bmatrix} \cos\omega_i & \sin\omega_i \\ -\sin\omega_i & \cos\omega_i \end{bmatrix},$$

with

$$\omega_i = \frac{2\pi i}{p} \tag{2.14}$$

and where p is the period of the seasonality component. Additionalharmonics can be added by composing the observation and system matrices, as follows.

Two distinct cases must be considered, respectively whether the period p is odd or even. When p is odd, we consider a Fourier seasonal representation of period p with m

harmonics with observation and system matrices as

$$\begin{aligned}
\mathsf{F} &= [1 \quad 0 \quad 1 \quad 0 \quad \cdots \quad 1 \quad 0]^T_{2p-2} \\
\mathsf{G} &= \text{block diag}\left(\mathsf{G}_1, \mathsf{G}_2, \ldots, \mathsf{G}_m\right) \\
\mathsf{G}_j &= \begin{bmatrix} \cos\omega_j & \sin\omega_j \\ -\sin\omega_j & \cos\omega_j \end{bmatrix} \\
\omega_j &= \frac{2\pi j}{p}, \qquad j = 1, 2, \ldots, m,
\end{aligned}$$

with $p = 2m - 1$. Whenever p is even, we add to the above specification an additional 1 to F and an extra diagonal element[3] to G of $\cos(\omega_m) = \cos(\pi) = -1$, that is

$$\begin{aligned}
\mathsf{F} &= [1 \quad 0 \quad 1 \quad 0 \quad \cdots \quad 1 \quad 0 \quad 1]^T_{2p-1} \\
\mathsf{G} &= \text{block diag}\left(\mathsf{G}_1, \mathsf{G}_2, \cdots, \mathsf{G}_{m-1}, 1\right) \\
\mathsf{G}_j &= \begin{bmatrix} \cos\omega_j & \sin\omega_j \\ -\sin\omega_j & \cos\omega_j \end{bmatrix} \\
\omega_j &= \frac{2\pi j}{p}, \qquad j = 1, 2, \ldots, m-1,
\end{aligned}$$

with $p = 2m$.

A specific case of the Fourier representation occurs when $\omega = \pi$, called the *Nyquist frequency* for which

$$\mathsf{F} = [1] \qquad \mathsf{G} = [-1]$$

We will refer to a Fourier component of period p with h harmonics as $\mathcal{F}(p, h)$.

Example. A Fourier seasonal component with $h = 2$ harmonics. An $\mathcal{F}(p, 2)$ component, where $\omega = 2\pi/p$:

$$\mathsf{F} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \qquad \mathsf{G} = \begin{bmatrix} \cos\omega & \sin\omega & 0 & 0 \\ -\sin\omega & \cos\omega & 0 & 0 \\ 0 & 0 & \cos 2\omega & \sin 2\omega \\ 0 & 0 & -\sin 2\omega & \cos 2\omega \end{bmatrix} \tag{2.15}$$

It is clear that this representation of seasonality has clear advantages over a full seasonal

[3] This corresponds to the *Nyquist frequency*.

representation in the context of our work. Since we are interested in real-time inference, lowering the computational burden by reducing the model size is of utmost importance. Using a full seasonal representation of p seasons will always incur a model dimension of size at least p. By using a Fourier representation we can reduce the model's seasonal component size to $2m$, the number of harmonics, which could be as low as $m = 1$. Additional harmonics can be added to capture more complex variation of that particular seasonal component.

2.2.3 Superposition

One of the advantages of this formulation is the possibility of constructing complex models from simple ones (*superposition*) or extracting components from a model (*decomposition*). If we consider $n > 1$ models with respective observation and system matrices F_i and G_i, $i = 1, \ldots, n$, in the univariate observation case which we are considering, the superimposed model can be constructed with the following rules:

- The observation matrix is the concatenation of the individual F_i matrices, for $i = 1, \ldots n$, that is

$$\mathsf{F} = [\mathsf{F}_1 \ \mathsf{F}_2 \ \cdots \ \mathsf{F}_n \ ]^T$$

- The state vector is the concatenation of the individual $\boldsymbol{\theta}_{i,t}$ such that

$$\boldsymbol{\theta}_t = (\boldsymbol{\theta}_{1,t} \ \boldsymbol{\theta}_{2,t} \ \cdots \ \boldsymbol{\theta}_{n,t})^T$$

- The state matrix is a block diagonal with:

$$\begin{aligned} \mathsf{G} &= \text{block diag}\,(\mathsf{G}_1, \mathsf{G}_2, \ldots, \mathsf{G}_n) \\ &= \begin{bmatrix} \mathsf{G}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathsf{G}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathsf{G}_n \end{bmatrix} \end{aligned}$$

- The system covariance matrix, W, is also obtained by the block diagonal:

$$\begin{aligned}\mathsf{W} &= \text{block diag}\left(\mathsf{W}_1, \mathsf{W}_2, \ldots, \mathsf{W}_n\right) \\ &= \begin{bmatrix} \mathsf{W}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathsf{W}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathsf{W}_n \end{bmatrix}\end{aligned}$$

In the specific case of the Normal DLM we know that the models can be combined linearly, by virtue of being precisely linear models.

In this case we have an additional rule, regarding the observation variance V. The superimposed observation variance can be obtained as the sum of the individual components' V_i:

$$V = \sum_{i=1}^{n} V_i$$

Example. Composed Normal DLM. If we assume two of the basic components referenced previously, respectively the $\mathcal{P}(2)$ (Section 2.2.1.2) and Fourier seasonal (Section 2.2.2.2) components we can combine them using superposition for a Normal DLM. If we consider the locally linear component as

$$\begin{aligned}\mathsf{W}_{linear} &= \begin{bmatrix} \tau_{l,1}^2 & 0 \\ 0 & \tau_{l,2}^2 \end{bmatrix} \\ V_{linear} &= \sigma_l^2 \\ \mathsf{F}_{linear} &= \begin{bmatrix} 1 & 0 \end{bmatrix}^T \\ \mathsf{G}_{linear} &= \mathbf{J}_2(1) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},\end{aligned}$$

and the seasonal component, with an odd numbered period p with $h = 2$ harmonics, and

ω_i as defined in (2.14), as

$$\begin{aligned}
\mathsf{W}_{seasonal} &= \begin{bmatrix} \tau^2_{s1,1} & 0 & 0 & 0 \\ 0 & \tau^2_{s1,2} & 0 & 0 \\ 0 & 0 & \tau^2_{s2,1} & 0 \\ 0 & 0 & 0 & \tau^2_{s2,2} \end{bmatrix} \\
V_{seasonal} &= \sigma^2_s \\
\mathsf{F}_{seasonal} &= [1 \quad 0 \quad 1 \quad 0]^T \\
\mathsf{G}_{seasonal} &= \begin{bmatrix} \cos\omega_1 & \sin\omega_1 & 0 & 0 \\ -\sin\omega_1 & \cos\omega_1 & 0 & 0 \\ 0 & 0 & \cos\omega_2 & \sin\omega_2 \\ 0 & 0 & -\sin\omega_2 & \cos\omega_2 \end{bmatrix}.
\end{aligned}$$

The composed model will then consist of

$$\mathsf{W} = \begin{bmatrix} \mathsf{W}_{linear} & \mathbf{0} \\ \mathbf{0} & \mathsf{W}_{seasonal} \end{bmatrix} = \begin{bmatrix} \tau_{l,1}^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \tau_{l,2}^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \tau_{s1,1}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \tau_{s1,2}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \tau_{s2,1}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \tau_{s2,2}^2 \end{bmatrix}$$

$$V = V_{linear} + V_{seasonal} = \sigma_l^2 + \sigma_s^2$$

$$\mathsf{F} = \begin{bmatrix} \mathsf{F}_{linear} \\ \mathsf{F}_{seasonal} \end{bmatrix} = [1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0]^T$$

$$\mathsf{G} = \begin{bmatrix} \mathsf{G}_{linear} & \mathbf{0} \\ \mathbf{0} & \mathsf{G}_{seasonal} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cos\omega_1 & \sin\omega_1 & 0 & 0 \\ 0 & 0 & -\sin\omega_1 & \cos\omega_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cos\omega_2 & \sin\omega_2 \\ 0 & 0 & 0 & 0 & -\sin\omega_2 & \cos\omega_2 \end{bmatrix}.$$

2.3 Dynamic Generalised Linear Models

2.3.1 Normal DLM

The Normal DLM (NDLM) is a specific case of the DGLM class where we have a linear observation model. This class of models is typically used in conjunction with continuous data. The observation equation (2.17) takes the form

$$p\left(y_t|\eta_t, \boldsymbol{\Phi}\right) = \frac{1}{\sqrt{2\pi V}} \exp\left\{-\frac{\left(y_t - \eta_t\right)^2}{2V}\right\}, \qquad y_t > -\infty, \eta_t < \infty, V > 0. \tag{2.16}$$

The state model follows the general form of the DGLM, that is

$$y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi} \sim \mathcal{N}\left(\mathsf{F}^T\boldsymbol{\theta}_t, V\right) \tag{2.17}$$

$$\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi} \sim \mathcal{N}\left(\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}\right) \tag{2.18}$$

The NDLM, usually called simply Dynamic Linear Model (DLM) (West & Harrison, 1997), is an important instance of DGLMs since common inference problems, such as state estimation and forecasting have an analytical solution (specifically the Kalman filter recursions as detailed in Chapter 4).

Taking into account the canonical formulation of the exponential family from (2.5), we can see from (2.16) that the corresponding quantities are

$$\begin{aligned} z\left(y_t\right) &= y_t \\ a\left(\phi_t\right) &= \phi^{-1} = V \\ b\left(\eta_t\right) &= \frac{\eta_t^2}{2} \\ c\left(y_t, \phi_t\right) &= \frac{1}{\sqrt{2\pi V}} \exp\left(-\frac{y_t^2}{2V}\right). \end{aligned}$$

In this case the link function (2.9) will be the *identity function*, that is

$$g\left(\mu_t\right) = \mu_t = \mathsf{F}^T\boldsymbol{\theta}_t$$

Example. Locally constant model

As specified in Section 2.2.1.1, for a $\mathcal{M} = \{\mathcal{P}\left(1\right)\}$ we will assume the following state and observation matrices

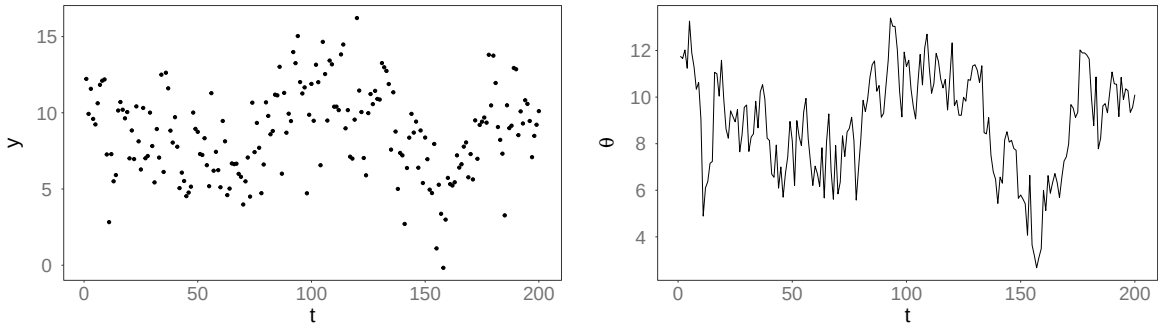
$$\mathsf{F} = \mathsf{G} = [1].$$

When used in conjunction with a linear observation model we have the following full specification

$$\begin{aligned} y_t &= \theta_t + \nu_t, \qquad \nu_t \sim \mathcal{N}\left(0, \sigma^2\right) \\ \theta_t &= \theta_{t-1} + \omega_t, \qquad \omega_t \sim \mathcal{N}\left(0, \tau^2\right). \end{aligned}$$

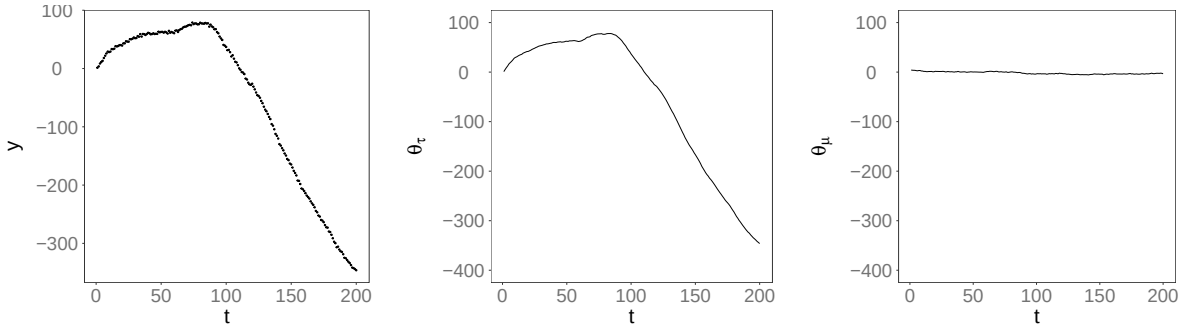
A realisation of a $\mathcal{M} = \{\mathcal{P}\left(1\right)\}$ NDLM with $\boldsymbol{\Phi} = \left\{\tau^2, \sigma^2\right\} = \{1.5, 3.0\}$ can be viewed in Figure 2.3 on the following page. This basically amounts to a random walk where the state θ_t corresponds to an underlying mean.

Example. Locally linear model



(a) $y_{1:t}$ (b) $\theta_{0:t}$

Figure 2.3: Realisation of a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with a state prior of $\theta_0 \sim \mathcal{N}(0, 100)$, $V = \sigma^2 = 3.0$ and $W = \tau^2 = 1.5$. Observations on the left and latent state of the right.



(a) $y_{1:t}$ (b) $\theta_{\tau,0:t}$ (c) $\theta_{\mu,0:t}$

Figure 2.4: Realisation of a $\mathcal{M} = \{\mathcal{P}(2)\}$ NDLM with a state prior of $\boldsymbol{\theta}_0 \sim \mathcal{N}\left((0,5)^T, 100\mathbf{I}\right)$, $V_t = \sigma^2 = 2.0$ and $\text{diag}(\mathsf{W}) = (0.1, 0.25)$. Observations $y_{1:t}$ (left-most) and latent state components ($\theta_{\tau,0:t}, \theta_{\mu,0:t}$, right).

A locally linear model, $\mathcal{M} = \{\mathcal{P}(2)\}$, consists of the following state and observation matrices

$$\mathsf{F} = \begin{bmatrix} 1 & 0 \end{bmatrix}^T, \qquad \mathsf{G} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

In this case, the state vector contains two components $\boldsymbol{\theta} = (\theta_\tau, \theta_\mu)^T$, which can be viewed as a series' mean and trend respectively. A realisation of this model, including the state components, can be viewed in Figure 2.4.

2.3.2 Poisson DLM

In the Poisson DLM (PoDLM) the observational model will follow a Poisson distribution and is usually applied when dealing with discrete data, such as count data. The observation model will be such that:

$$p\left(y_t|\eta_t\right) = e^{-e^{\eta_t}} \frac{e^{\eta_t y_t}}{y_t!}, \qquad \eta_t > 0. \tag{2.19}$$

From (2.19) we verify that the canonical quantities of (2.5) are

$$\begin{aligned} z\left(y_t\right) &= y_t \\ a\left(\phi_t\right) &= \phi_t = 1 \\ b\left(\eta_t\right) &= e^{\eta_t} \\ c\left(y_t, \phi_t\right) &= \log\left(\frac{1}{y_t!}\right), \end{aligned}$$

where $\eta_t = \mathsf{F}^T\boldsymbol{\theta}_t$. It follows that, according to (2.6) and (2.7)

$$\mathrm{E}\left[z\left(y_t\right)|\eta_t\right] = \mathrm{E}\left[y_t|\eta_t\right] = b'\left(\eta_t\right) = e^{\eta_t} \tag{2.20}$$

$$\mathrm{Var}\left[z\left(y_t\right)|\eta_t\right] = \mathrm{Var}\left[y_t|\eta_t\right] = a\left(\phi_t\right) b''\left(\eta_t\right) = e^{\eta_t}. \tag{2.21}$$

The full form of the Poisson DLM will then be:

$$\begin{aligned} y_t|\eta_t &\sim \mathrm{Po}\left(e^{\eta_t}\right) \\ \eta_t &= \mathsf{F}^T\boldsymbol{\theta}_t. \end{aligned} \tag{2.22}$$

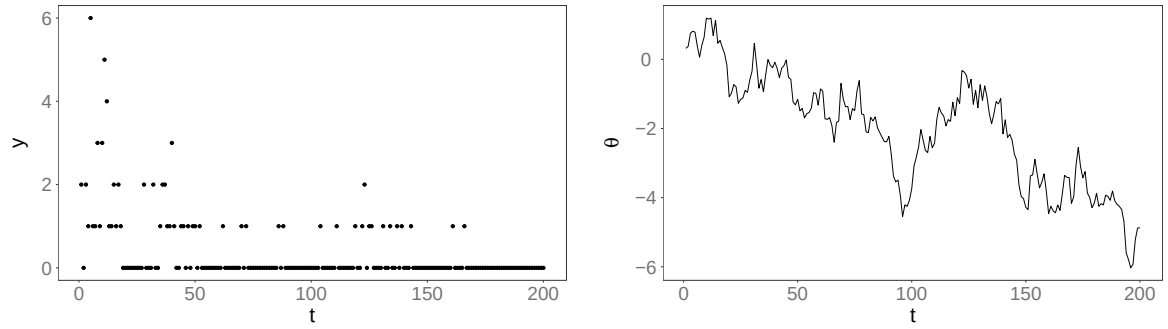
The state model will follow the standard DGLM form (2.3). As it is clear from the above definition, the parameter set for the PoDLM will simply be state transition variance, that is $\boldsymbol{\Phi} = \{\mathsf{W}\}$.

Example. Locally constant Poisson DLM

A $\mathcal{M} = \{\mathcal{P}\left(1\right)\}$ PoDLM with $\theta_0 \sim \mathcal{N}\left(2, 10\right)$ and $\tau^2 = 0.15$ will take the form

$$\begin{aligned} y_t|\theta_t &\sim \mathrm{Po}\left(e^{\theta_t}\right) \\ \theta_t|\theta_{t-1} &\sim \mathcal{N}\left(\theta_{t-1}, 0.15\right). \end{aligned}$$

A realisation of this model, along with the state can be viewed in Figure 2.5.



(a) $y_{1:t}$ (b) $\theta_{0:t}$

Figure 2.5: Observations (*left*) and latent states (*right*) for a realisation of a $\mathcal{M} = \{\mathcal{P}(1)\}$ PoDLM with $\boldsymbol{\Phi} = \{W\} = \{0.15\}$.

2.3.3 Binomial DLM

The Binomial DLM (BDLM) is a another special case of DGLMs where the state evolution will have the form as (2.18), but the observation model will take the form of a Binomial distribution, that is

$$p(y_t|p_t) = \binom{n_t}{y_t} p_t^{y_t} (1-p_t)^{n_t - y_t}, \qquad n_t = 1, 2, \ldots, \quad y_t = 0, 1, \ldots, n_t, \quad 0 < p_t < 1.$$

Looking at the DGLM canonical form in (2.5), we can see that

$$\begin{aligned} z(y_t) &= \frac{y_t}{n_t} \\ \eta_t &= \log\left(\frac{p_t}{1-p_t}\right) \\ a(\phi_t) &= \phi_t^{-1} = n_t^{-1} \\ b(\eta_t) &= \log(1 + e^{\eta_t}) \\ c(y_t, \phi_t) &= \log\binom{n_t}{y_t}. \end{aligned}$$

As in the previous sections, we can calculate the DGLM mean and variance using (2.6) and (2.7). In this case we have

$$\begin{aligned} \mathrm{E}[z(y_t)|\eta_t] &= b'(\eta_t) = \mathrm{logit}^{-1}\{\eta_t\} \\ \mathrm{Var}[z(y_t)|\eta_t] &= a(\phi_t)\, b''(\eta_t) = n_t^{-1} \frac{e^{\eta_t}}{(e^{\eta_t}+1)^2}, \end{aligned}$$

therefore

$$\begin{aligned}
\mathrm{E}\left[y_t|\eta_t\right] &= n_t \mathrm{logit}^{-1}\left\{\eta_t\right\} \\
&= n_t \frac{\exp\left\{\eta_t\right\}}{\exp\left\{\eta_t\right\}+1} \\
&= n_t \frac{\exp\left\{\log\left(p_t\left(1-p_t\right)^{-1}\right)\right\}}{\exp\left\{\log\left(p_t\left(1-p_t\right)^{-1}\right)\right\}+1} \\
&= n_t \frac{p_t}{p_t+\left(1-p_t\right)} \\
&= n_t p_t \qquad (2.23) \\
\mathrm{Var}\left[y_t|\eta_t\right] &= n_t \frac{\exp\left\{\eta_t\right\}}{\left(\exp\left\{\eta_t\right\}+1\right)^2} \\
&= n_t \frac{\exp\left\{\log\left(p_t\left(1-p_t\right)^{-1}\right)\right\}}{\left[\exp\left\{\log\left(p_t\left(1-p_t\right)^{-1}\right)\right\}+1\right]^2} \\
&= n_t \frac{p_t\left(1-p_t\right)^{-1}}{\left[p_t\left(1-p_t\right)^{-1}+1\right]^2} \\
&= n_t p_t\left(1-p_t\right). \qquad (2.24)
\end{aligned}$$

The full form of the Binomial DLM will then be:

$$\begin{aligned}
y_t|\eta_t &\sim \mathrm{Binom}\left(\mathrm{logit}^{-1}\{\eta_t\}, n_t\right) \qquad (2.25) \\
\eta_t &= \mathsf{F}^T\boldsymbol{\theta}_t \\
\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}\right).
\end{aligned}$$

Where the $\mathrm{logit}\left(\cdot\right)$ function and its inverse, $\mathrm{logit}^{-1}\left(\cdot\right)$, are respectively

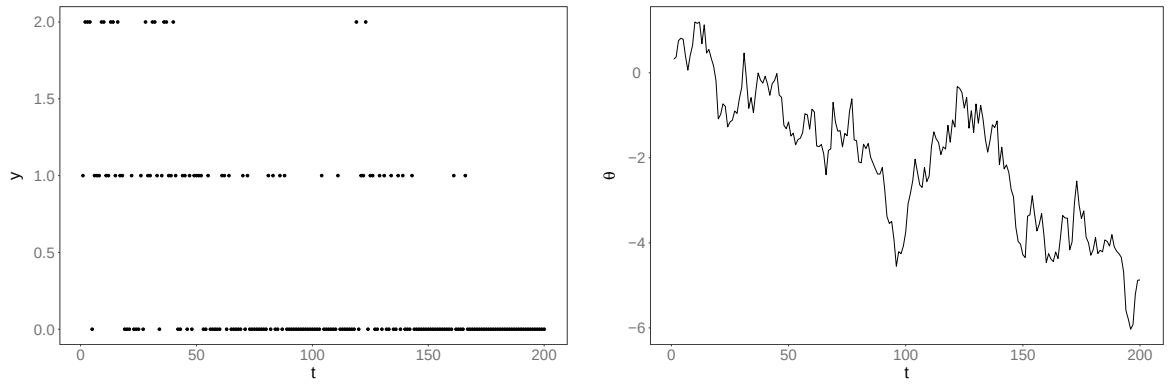
$$\mathrm{logit}\left(x\right) = \log\left(\frac{x}{1-x}\right) \qquad (2.26)$$

$$\mathrm{logit}^{-1}\left(x\right) = \frac{\exp\left\{x\right\}}{\exp\left\{x\right\}+1}. \qquad (2.27)$$

In this thesis we will always assume n_t to be known and it will commonly take the value $n_t = 1$ to model the binary nature of our datasets.

Example. Locally constant Binomial DLM.

In Figure 2.6 we can see a realisation of a $\mathcal{P}\left(1\right)$ BDLM with parameter set $\boldsymbol{\Phi} = \{W\} = 0.15$ and state prior $\theta_0 \sim \mathcal{N}\left(2, 10\right)$ and $n = 3$.



(a) $y_{1:t}$ (b) $\theta_{0:t}$

Figure 2.6: Observations (*left*) and latent states (*right*) for a realisation of a $\mathcal{M} = \{\mathcal{P}(1)\}$ PoDLM with a state prior of $\theta_0 \sim \mathcal{N}(0, 100)$,$V = \sigma^2 = 3.0, W = \tau^2 = 1.5$ and $n = 3$.

2.3.4 Summary

A summary for the values of the natural parameter, η_t, and $b(\eta_t)$ for each of the considered DGLMs is presented in table 2.1.

Observation		η	$b(\eta)$
Gaussian	$\mathcal{N}(\mu, V)$	μ	η
Poisson	$\text{Po}(\mu)$	$\log \mu$	$\exp(\eta)$
Binomial	$\text{Binom}(n, p)$	$\text{logit}(p)$	$n \log(1 + \exp \eta)$

Table 2.1: Summary of η_t and $b(\eta_t)$ for different DGLMs

2.4 Inference

As mentioned in Chapter 1 our objective is to perform inference in DGLMs. By inference we include

- Online state estimation (when the model parameters, $\boldsymbol{\Phi}$, are known)
- Online state and parameter estimation
- One-step ahead forecast, both for state and observations
- k-step ahead forecasts
- Model assessment

2.4.1 State estimation

The problem of pure state estimation can divided into two approaches, namely *online* and *offline* estimation. Online estimation, also commonly known as *filtering* amounts to establishing the state $\boldsymbol{\theta}_t$ given the data $\mathcal{D}_t = \{y_1, \ldots, y_t\}$ for $t = 1, \ldots, T$. For pure state estimation we assume that model's parameter set $\boldsymbol{\Phi}$ is known (and as such will be dropped from the notation). That is, we are trying to estimate

$$p\left(\boldsymbol{\theta}_{0:t}|\mathcal{D}_t\right),$$

or

$$p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right).$$

Regarding offline state estimation, that is, given the totality of the observations $\mathcal{D}_T = \{y_1, \ldots, y_T\}$, estimating

$$p\left(\boldsymbol{\theta}_{0:T}|\mathcal{D}_T\right),$$

is usually referred in the literature as *smoothing*. As we will see in Sections 4 and 11, the problem is greatly simplified for the Normal DLM case, where we can obtain analytic expressions for this purpose. This is not the case, however, for other DGLMs.

2.4.2 Forecasting

If at time t we have the state vector estimation $p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right)$, the k-step ahead *observation forecast*, for $k > 0$, can be defined as

$$p\left(y_{t+k}|\mathcal{D}_t\right) = \int p\left(y_{t+k}|\boldsymbol{\theta}_{t+k}\right) p\left(\boldsymbol{\theta}_{t+k}|\mathcal{D}_t\right) d\boldsymbol{\theta}_{t+k} \tag{2.28}$$

As for state forecast, we are interested in

$$p\left(\boldsymbol{\theta}_{t+k}|\mathcal{D}_t\right).$$

As with state estimation, analytical expressions of the forecast are available for the Normal DLM, but not for non-linear models. Additional approximation methods will be discussed to perform state and observation forecasts in non-linear DGLMs. A special case in forecast is the one of $k = 1$ called the *one-step ahead* forecast.

2.4.3 Anomaly detection

Another advantage of the DGLM, especially in the context of real-time streaming data, is the natural way in which model diagnostics and anomaly detection can be performed in real-time as observations arrive. One of the employed methods is the calculation of a *discrepancy* value. This method consists of quantifying the discrepancy, effectively a Pearson type residual, at each time t, between the observed data, y_t and the observation prediction $\hat{y}_t$. The threshold can be specified as a function of the predicted value and standard deviation of $\hat{y}_t$. To perform a *discrepancy* test against the observed data, y_t, we calculate

$$\begin{aligned} \mathrm{E}\left[y_t\right] &= \mathrm{E}_{\boldsymbol{\theta}_t}\left[\mathrm{E}\left[y_t|\boldsymbol{\theta}_t\right]\right] \\ \mathrm{Var}\left[y_t\right] &= \mathrm{E}_{\boldsymbol{\theta}_t}\left[\mathrm{Var}\left[y_t|\boldsymbol{\theta}_t\right]\right] + \mathrm{Var}_{\boldsymbol{\theta}_t}\left[\mathrm{E}\left[y_t|\boldsymbol{\theta}_t\right]\right] \end{aligned}$$

The discrepancy, $d\left(\cdot\right)$, will then be

$$d\left(y_t\right) = \frac{\left|y_t - \mathrm{E}\left[y_t\right]\right|}{\sqrt{\mathrm{Var}\left[y_t\right]}} \tag{2.29}$$

As we've seen is Section 2.1, we can derive $\mathrm{E}\left[z\left(y_t\right)|\eta_t\right]$ and $\mathrm{Var}\left[z\left(y_t\right)|\eta_t\right]$ for a DGLM from the exponential family observational model canonical form (2.5) as

$$\begin{aligned} \mathrm{E}\left[z\left(y_t\right)|\eta_t\right] &= \frac{db\left(\eta_t\right)}{d\eta_t}, \\ \mathrm{Var}\left[z\left(y_t\right)|\eta_t\right] &= a\left(\phi\right)\frac{d^2b\left(\eta_t\right)}{d\eta_t}, \end{aligned}$$

assuming our previous formulation of a DGLM in (2.5). In the following sections we will elaborate the discrepancy form for each of the considered observational models with Table 2.2 summarising the results. In the experimental results a threshold of 3 standard deviations was used to flag possible outliers.

2.4.3.1 Normal DLM

For the NDLM since we have

$$\begin{aligned} a\left(\phi_t\right) &= \phi_t^{-1} = V \\ \eta_t &= \mu_t = \mathsf{F}^T\boldsymbol{\theta}_t \\ b\left(\eta_t\right) &= \frac{\eta_t^2}{2}, \end{aligned}$$

it follows that

$$\begin{aligned}\mathrm{E}\left[y_t|\eta_t\right] &= \frac{d\left(\frac{\eta_t^2}{2}\right)}{d\eta_t} = \eta_t = \mathsf{F}^T\boldsymbol{\theta}_t \\ \mathrm{Var}\left[y_t|\eta_t\right] &= \underbrace{a\left(\phi\right)}_{V}\underbrace{\frac{d^2b\left(\eta_t\right)}{d\eta_t}}_{1} = V.\end{aligned}$$

The discrepancy will take then the form expected for a DLM, that is

$$d\left(y_t\right) = \frac{\left|y_t - \mathsf{F}^T\boldsymbol{\theta}_t\right|}{\sqrt{V}}. \tag{2.30}$$

2.4.3.2 Poisson DLM

For the PoDLM, according to Section 2.3.2, we have

$$\begin{aligned}a\left(\phi_t\right) &= \phi_t = 1 \\ \eta_t &= \mathsf{F}^T\boldsymbol{\theta}_t \\ b\left(\eta_t\right) &= -e^{\eta_t}.\end{aligned}$$

Using the derivations in (2.20) and (2.21), we can write

$$\begin{aligned}\mathrm{E}\left[y_t|\eta_t\right] &= \exp\left\{\eta_t\right\} = \exp\left\{\mathsf{F}^T\boldsymbol{\theta}_t\right\} \\ \mathrm{Var}\left[y_t|\eta_t\right] &= \exp\left\{\eta_t\right\} = \exp\left\{\mathsf{F}^T\boldsymbol{\theta}_t\right\},\end{aligned}$$

resulting in a discrepancy form of

$$d\left(y_t\right) = \frac{\left|y_t - \exp\left\{\mathsf{F}^T\boldsymbol{\theta}_t\right\}\right|}{\sqrt{\exp\left\{\mathsf{F}^T\boldsymbol{\theta}_t\right\}}}. \tag{2.31}$$

2.4.3.3 Binomial DLM

For the Binomial DLM, using Section 2.3.3, we have

$$\begin{aligned}a\left(\phi_t\right) &= \phi_t^{-1} = n_t^{-1} \\ \eta_t &= \log\left\{\frac{p_t}{1-p_t}\right\} \\ b\left(\eta_t\right) &= \log\left\{1 + \exp\eta_t\right\}.\end{aligned}$$

Observation		$d(y_t)$
Gaussian	$\mathcal{N}\left(\mathsf{F}^T\boldsymbol{\theta}_t, V\right)$	$\left\|y_t - \mathsf{F}^T\boldsymbol{\theta}_t\right\| V^{-1/2}$
Poisson	$\mathrm{Po}\left(\exp\left\{\mathsf{F}^T\boldsymbol{\theta}_t\right\}\right)$	$\left\|y_t - \exp\left\{\mathsf{F}^T\boldsymbol{\theta}_t\right\}\right\| \exp\left\{\mathsf{F}^T\boldsymbol{\theta}_t\right\}^{-1/2}$
Binomial	$\mathrm{Binom}(n, p)$	$\left\|y_t - n_t p_t\right\| \left[n_t p_t (1-p_t)\right]^{-1/2}$

Table 2.2: Discrepancy ($d(y_t)$) calculation for different classes of DGLMs

As mentioned in Section 2.3.3 on page 18, in this thesis we will always consider n_t to be fixed and known. We have seen that according to (2.23) and (2.24) we can write

$$\begin{aligned} \mathrm{E}\left[y_t|\eta_t\right] &= n_t p_t \\ \mathrm{Var}\left[y_t|\eta_t\right] &= n_t p_t (1-p_t). \end{aligned}$$

Consequently the discrepancy will be

$$d(y_t) = \frac{|y_t - n_t p_t|}{\sqrt{n_t p_t (1-p_t)}}. \tag{2.32}$$

Part II

Online State Estimation

Chapter 3

Bayesian Filtering

In order to perform inference in DGLMs, the main objective is to estimate the unobserved sequence of states $\boldsymbol{\theta}_{0:t} = \{\boldsymbol{\theta}_0, \dots, \boldsymbol{\theta}_t\}$ and the parameter set $\boldsymbol{\Phi} = \{\Phi_1, \dots, \Phi_n\}$ given the observed data, $\mathcal{D}_t = \{y_1, \dots, y_t\}$. That is, we are trying to estimate the joint density

$$p\left(\boldsymbol{\theta}_{0:t}, \boldsymbol{\Phi} | \mathcal{D}_t\right). \tag{3.1}$$

We will first look at some methods to estimate the state vectors in an online fashion, that is estimating $\boldsymbol{\theta}_t$ using $\mathcal{D}_t$ sequentially (with $t = 1, 2, 3, \dots$) while considering the parameters known. These methods will provide the fundamental framework from which extensions can be used to simultaneously estimate states and parameters in Chapter 7.

As mentioned previously, assuming the model's parameters to be known (on account of which we will omit the dependency of $\boldsymbol{\Phi}$ for the time being), in DGLMs the problem of estimating the unobserved state vectors $\boldsymbol{\theta}_{0:t}$ can be expressed as

$$p\left(\boldsymbol{\theta}_{0:t} | \mathcal{D}_t\right) = \frac{p\left(\boldsymbol{\theta}_{0:t}, \mathcal{D}_t\right)}{p\left(\mathcal{D}_t\right)}, \tag{3.2}$$

where

$$\begin{aligned} p\left(\boldsymbol{\theta}_{0:t}, \mathcal{D}_t\right) &= p\left(\mathcal{D}_t | \boldsymbol{\theta}_{0:t}\right) p\left(\boldsymbol{\theta}_{0:t}\right), \\ p\left(\mathcal{D}_t\right) &= \int p\left(\boldsymbol{\theta}_{0:t}, \mathcal{D}_t\right) d\boldsymbol{\theta}_{0:t}. \end{aligned} \tag{3.3}$$

The Markovian nature of the DGLMs can, however, be exploited to provide a recursive formulation for the state estimation in (3.2). This is crucial for allowing *online* inference in DGLMs, since it provides us with a tool to perform computations for each time step t independently from the previous time steps. If we denote $\mathcal{D}_{t-1} = \{y_1, y_2, \dots, y_{t-1}\}$, the

full state posterior (3.2) can then be re-arranged as a recursive update.

Using Bayes' theorem[1], and using the property that

$$p\left(\mathcal{D}_t\right) = p\left(y_t, \mathcal{D}_{t-1}\right), \tag{3.4}$$

we can write $p\left(\boldsymbol{\theta}_{0:t}|\mathcal{D}_t\right)$ as

$$\begin{aligned} p\left(\boldsymbol{\theta}_{0:t}|\mathcal{D}_t\right) &= \frac{p\left(y_t, \mathcal{D}_{t-1}, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{0:t-1}\right)}{p\left(y_t|\mathcal{D}_{t-1}\right) p\left(\mathcal{D}_{t-1}\right)} \\ &= \frac{p\left(y_t|\mathcal{D}_{t-1}, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{0:t-1}\right) p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_{t-1}\right) p\left(\boldsymbol{\theta}_{0:t-1}|\mathcal{D}_{t-1}\right) p\left(\mathcal{D}_{t-1}\right)}{p\left(y_t|\mathcal{D}_{t-1}\right) p\left(\mathcal{D}_{t-1}\right)} \end{aligned} \tag{3.5}$$

Applying once more one of the properties arising from the Markovian nature of the DGLMs, this time (2.11), that is

$$p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_{t-1}\right) = p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\right),$$

we can simplify (3.5) into

$$p\left(\boldsymbol{\theta}_{0:t}|\mathcal{D}_t\right) = \frac{p\left(y_t|\boldsymbol{\theta}_t\right) p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\right)}{p\left(y_t|\mathcal{D}_{t-1}\right)} p\left(\boldsymbol{\theta}_{0:t-1}|\mathcal{D}_{t-1}\right) \tag{3.6}$$

where $p\left(y_t|\mathcal{D}_{t-1}\right)$ is a normalising constant taking the form

$$p\left(y_t|\mathcal{D}_{t-1}\right) = \int p\left(y_t|\boldsymbol{\theta}_t\right) p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\right) p\left(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}\right) d\boldsymbol{\theta}_{t-1:t}. \tag{3.7}$$

The posterior distribution can then be expressed recursively as

$$p\left(\boldsymbol{\theta}_{0:t}|\mathcal{D}_t\right) \propto p\left(\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_{t-1}\right) \underbrace{p\left(y_t|\boldsymbol{\theta}_t\right)}_{\text{measurement}} \underbrace{p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\right)}_{\text{system}} \tag{3.8}$$

However, to perform online state estimation we need to perform the estimation as the observations appear, *i.e.* we need to estimate the *current state* (conditional on the observations). This is usually referred in the literature as *Bayesian filtering* and targets the state's marginal posterior

$$p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right). \tag{3.9}$$

The goal is therefore to estimate $p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right)$ having, at time $t-1$, $p\left(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}\right)$. To do so, we can divide *filtering* into two separate stages, the *prediction* and the *update* steps.

[1] *cf.* Section F.2 on page 318.

Predict step In this step calculate the *predictive state density* given the observations up to time $t-1$. If we assume that the posterior at time $t-1$, $p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1})$, is known, then the joint distribution of $\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}$ and $\mathcal{D}_{t-1}$ can be calculated (using the property (2.11)) as:

$$\begin{aligned} p(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) &= \underbrace{p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathcal{D}_{t-1})}_{p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})} p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) \\ &= p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})\, p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) \end{aligned}$$

To obtain the state prediction density, that is the distribution of $\boldsymbol{\theta}_t$ given the previous observations $\mathcal{D}_{t-1}$ we use the Chapman-Kolmogorov[2] equation:

$$p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) = \int p(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1})\, d\boldsymbol{\theta}_{t-1} \tag{3.10}$$

$$= \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathcal{D}_{t-1})\, p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1})\, d\boldsymbol{\theta}_{t-1} \tag{3.11}$$

$$= \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})\, p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1})\, d\boldsymbol{\theta}_{t-1} \tag{3.12}$$

Update step Having the prior $p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})$ and the observation likelihood $p(y_t|\boldsymbol{\theta}_t)$ we can formulate the state's posterior by using Bayes' theorem and the Markovian property (2.13):

$$\begin{aligned} p(\boldsymbol{\theta}_t|\mathcal{D}_t) &= \frac{\overbrace{p(y_t|\boldsymbol{\theta}_t, \mathcal{D}_{t-1})}^{p(y_t|\boldsymbol{\theta}_t)} p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})}{p(y_t|\mathcal{D}_{t-1})} \\ &= \frac{p(y_t|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})}{p(y_t|\mathcal{D}_{t-1})} \end{aligned} \tag{3.13}$$

where the denominator is a normalisation constant equal to

$$p(y_t|\mathcal{D}_{t-1}) = \int p(y_t|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})\, d\boldsymbol{\theta}_t. \tag{3.14}$$

These theoretical results constitute the cornerstone of sequential Bayesian state estimation and in the following chapters we will look at some common methodologies which employ them to perform online filtering.

[2] *cf.* Section F.3.

Chapter 4

Kalman filter

In most DGLMs the estimation of $p(\boldsymbol{\theta}_{0:t}|\mathcal{D}_t)$ as described in (3.6) is not possible using analytical methods. The exception is the specific case of the NDLM where this solution exists. This method is widely known as the Kalman Filter (KF, Kalman (1960)).

The KF is a recursive method, meaning that only the estimated state from the previous time step and the current measurement are needed to compute the estimate for the current state (as such it is an *online estimation* method). It is also the optimal linear estimator, meaning that any other linear estimators will have a higher error covariance. It is a thoroughly researched method and, due to its optimality and simplicity, it is also widely applied to online state estimation in NDLMs in the case where the model's parameters are known.

If we again consider a NDLM such as presented in Section 2.3.1, the KF allows us to perform the state estimation $\hat{\boldsymbol{\theta}}_t$ given the observation at time t, y_t. Algorithmically, the KF can be separated into two distinct steps, as discussed in the previous chapter, the *prediction* and the *update* steps. We will use the notation $\hat{X}_{t|t}$ to denote the estimation of a quantity X at time t, given the observation y_t and $\hat{X}_{t|t-1}$ do denote the *predicted* value of quantity X at time t before y_t is taken into account.

To perform the derivation of the KF recursions, we will use the following result for standard Normal theory:

Considering a multivariate Normal density defined as

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \tag{4.1}$$
$$\boldsymbol{\mu} \in \mathbb{R}^n, \boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}, \mathbf{x} \in \boldsymbol{\mu} + \text{span}(\boldsymbol{\Sigma}) \subseteq \mathbb{R}^n,$$

if two random variables, $\mathbf{X}$ and $\mathbf{Y}$, have a joint Normal probability density

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix} \right), \tag{4.2}$$

then

$$\begin{aligned} \mathbf{X} &\sim \mathcal{N}(\mathbf{a}, \mathbf{A}) \\ \mathbf{Y} &\sim \mathcal{N}(\mathbf{b}, \mathbf{B}), \end{aligned}$$

and the conditional distribution $\mathbf{X}|\mathbf{Y}$ will be

$$\mathrm{E}[\mathbf{X}|\mathbf{y}] = \mathbf{a} + \mathbf{C}\mathbf{B}^{-1}(\mathbf{y} - \mathbf{b}) \tag{4.3}$$

$$\mathrm{Var}[\mathbf{X}|\mathbf{y}] = \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^T. \tag{4.4}$$

Predict step The *prediction step*, as described by (3.12), consists of calculating $p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})$. If we assume the NDLM to have state priors as defined in (2.4), that is $\boldsymbol{\theta}_0|\mathbf{m}_0, \mathsf{C}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0)$, then at each step t we can update the state vector according to

$$\boldsymbol{\theta}_t = \mathsf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim \mathcal{N}(\mathbf{0}, \mathsf{W}).$$

Since $\boldsymbol{\theta}_{t-1}$ and $\boldsymbol{\omega}_t$ are independent, we can write

$$\mathrm{E}[\boldsymbol{\theta}_t|\mathcal{D}_{t-1}] = \mathsf{G}\mathbf{m}_{t-1} = \mathbf{a}_t \tag{4.5}$$

$$\begin{aligned} \mathrm{Var}[\boldsymbol{\theta}_t|\mathcal{D}_{t-1}] &= \mathrm{E}\left[(\boldsymbol{\theta}_t - \mathbf{a}_t)(\boldsymbol{\theta}_t - \mathbf{a}_t)^T\right] \\ &= \mathsf{G}\mathsf{C}_{t-1|t-1}\mathsf{G}^T + \mathsf{W} \qquad (4.6) \\ &= \mathsf{R}_t \end{aligned}$$

We denote, at time $t-1$, $\mathbf{m}_{t-1}$ and $\mathsf{C}_{t-1|t-1}$ respectively as the KF's *first moment* and *covariance* and label (4.5) and (4.6) respectively as R_t and $\mathbf{a}_t$. We can then define the state's *propagation density* as

$$\boldsymbol{\theta}_t|\mathcal{D}_{t-1} \sim \mathcal{N}(\mathbf{a}_t, \mathsf{R}_t). \tag{4.7}$$

If we now consider that the observation equation can be written as

$$y_t = \mathsf{F}^T \boldsymbol{\theta}_t + \nu_t, \qquad \nu_t \sim \mathcal{N}(0, V), \tag{4.8}$$

again, by the assumption that $\boldsymbol{\theta}_t$ and ν_t are independent, we know that y_t will be normally distributed. We can then write

$$\begin{aligned}
\mathrm{E}[y_t|\mathcal{D}_{t-1}] &= \mathsf{F}^T \mathbf{a}_t = f_t \\
\mathrm{Var}[y_t|\mathcal{D}_{t-1}] &= \mathrm{E}\left[(y_t - f_t)(y_t - f_t)^T\right] \\
&= \mathrm{E}\left[\left(\underbrace{\mathsf{F}^T\boldsymbol{\theta}_t + \nu_t}_{y_t} - \underbrace{\mathsf{F}^T\mathbf{a}_t}_{f_t}\right)\left(\mathsf{F}^T\boldsymbol{\theta}_t + \nu_t - \mathsf{F}^T\mathbf{a}_t\right)^T\right] \\
&= \mathrm{E}\left[\left\{\mathsf{F}^T(\boldsymbol{\theta}_t - \mathbf{a}_t) + \nu_t\right\}\left\{\mathsf{F}^T(\boldsymbol{\theta}_t - \mathbf{a}_t) + \nu_t\right\}^T\right] \\
&= \mathsf{F}^T\mathsf{R}_t\mathsf{F} + V = Q_t
\end{aligned}$$

which is usually referred to as the *Kalman's predictive density*, the prediction of y at $t-1$:

$$y_t|\mathcal{D}_{t-1} \sim \mathcal{N}(f_t, Q_t). \tag{4.9}$$

The calculation of these two quantities, (4.7) and (4.9), concludes the *prediction step*.

Update step To perform the *update step*, we need to find the posterior of $\boldsymbol{\theta}_t$ given the observed data, that is $p(\boldsymbol{\theta}_t|y_t)$. As defined in (3.13), we need to find

$$p(\boldsymbol{\theta}_t|\mathcal{D}_t) = \frac{p(y_t|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})}{p(y_t|\mathcal{D}_{t-1})}.$$

Since we are considering joint normality, according to the NDLM model specification, we simply need to construct the joint normal distribution $\begin{pmatrix} \boldsymbol{\theta}_t \\ y_t \end{pmatrix}$. We start by writing $\boldsymbol{\theta}_t$ and y_t as follows:

$$\begin{aligned}
\boldsymbol{\theta}_t &= \mathbf{a}_t + (\boldsymbol{\theta}_t - \mathbf{a}_t) \\
y_t &= \underbrace{f_t}_{\mathsf{F}^T\mathbf{a}_t} + \underbrace{y_t}_{\mathsf{F}^T\boldsymbol{\theta}_t + \nu_t} - f_t \\
&= \mathsf{F}^T\mathbf{a}_t + \mathsf{F}^T\boldsymbol{\theta}_t + \nu_t - \mathsf{F}^T\mathbf{a}_t \\
&= \mathsf{F}^T\mathbf{a}_t + \mathsf{F}^T(\boldsymbol{\theta}_t - \mathbf{a}_t) + \nu_t
\end{aligned}$$

The covariance of $\boldsymbol{\theta}_t$ and y_t can then be written as

$$\begin{aligned}
\operatorname{Cov}[\boldsymbol{\theta}_t, y_t] &= \mathrm{E}\left[(\boldsymbol{\theta}_t - a_t)\left(\underbrace{y_t}_{\mathsf{F}^T\boldsymbol{\theta}_t+\nu_t} - \underbrace{f_t}_{\mathsf{F}^T\mathbf{a}_t}\right)^T\right] \\
&= \mathrm{E}\left[(\boldsymbol{\theta}_t - \mathbf{a}_t)\left\{\mathsf{F}^T(\boldsymbol{\theta}_t - \mathbf{a}_t) + \nu_t\right\}^T\right] \\
&= \mathrm{E}\left[(\boldsymbol{\theta}_t - \mathbf{a}_t)\left\{(\boldsymbol{\theta}_t - \mathbf{a}_t)\mathsf{F} + \nu_t^T\right\}\right] \\
&= \mathrm{E}\left[(\boldsymbol{\theta}_t - \mathbf{a}_t)(\boldsymbol{\theta}_t - \mathbf{a}_t)\mathsf{F} + (\boldsymbol{\theta}_t - \mathbf{a}_t)\nu_t^T\right] \\
&= \underbrace{\mathrm{E}\left[(\boldsymbol{\theta}_t - \mathbf{a}_t)(\boldsymbol{\theta}_t - \mathbf{a}_t)\right]}_{\mathsf{R}_t}\mathsf{F} + \mathrm{E}\left[(\boldsymbol{\theta}_t - \mathbf{a}_t)\nu_t^T\right] \\
&= \mathsf{R}_t\mathsf{F}
\end{aligned}$$

We can then write the joint normal distribution in the form (4.2):

$$\begin{pmatrix}\boldsymbol{\theta}_t \\ y_t\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mathbf{a}_t \\ \mathsf{F}^T\mathbf{a}_t\end{pmatrix}, \begin{bmatrix}\mathsf{R}_t & \mathsf{R}_t\mathsf{F} \\ \mathsf{F}^T\mathsf{R}_t & \mathsf{F}^T\mathsf{R}_t\mathsf{F} + V\end{bmatrix}\right) \tag{4.10}$$

to write the result of

$$\boldsymbol{\theta}_t | y_t \sim \mathcal{N}(\mathbf{m}_t, \mathsf{C}_t).$$

By using the results of (4.3) and (4.4) we have for $\mathbf{m}_t$

$$\begin{aligned}
\mathbf{m}_t &= \mathrm{E}[\boldsymbol{\theta}_t | y_t] = \underbrace{\mathbf{a}_t}_{a} + \underbrace{\mathsf{R}_t\mathsf{F}}_{C}\underbrace{\left(\mathsf{F}^T\mathsf{R}_t\mathsf{F} + V\right)^{-1}}_{B^{-1}}\left(\underbrace{y_t}_{Y} - \underbrace{\mathsf{F}^T\mathbf{a}_t}_{b}\right) \\
&= \mathbf{a}_t + \mathsf{R}_t\mathsf{F}\left(\mathsf{F}^T\mathsf{R}_t\mathsf{F} + V\right)^{-1}(y - f_t),
\end{aligned} \tag{4.11}$$

and for C_t:

$$\begin{aligned}
\mathsf{C}_t &= \operatorname{Var}[\boldsymbol{\theta}_t | y_t] \\
&= \mathsf{R}_t - \mathsf{R}_t\mathsf{F}\left(\mathsf{F}^T\mathsf{R}_t\mathsf{F} + \mathsf{W}\right)^{-1}\mathsf{F}\mathsf{R}_t.
\end{aligned} \tag{4.12}$$

The notation can be simplified by introducing some quantities commonly used in the literature. We will denote

$$\begin{aligned}
\mathsf{A}_t &= \mathsf{R}_t\mathsf{F} \\
Q_t &= \mathsf{F}\mathsf{R}_t\mathsf{F}^T + V,
\end{aligned}$$

and we can define the *Kalman gain* as

$$\begin{aligned} \mathbf{K}_t &= \mathsf{R}_t \mathsf{F}^T \left(\mathsf{F}\mathsf{R}_t\mathsf{F}^T + V\right)^{-1} \\ &= \mathsf{A}_t Q_t^{-1}. \end{aligned}$$

Another useful quantity is usually referred to as the *Kalman error* and is defined as

$$e_t = y_t - f_t.$$

Applying substitutions, the final moments calculations become

$$\begin{aligned} \mathbf{m}_t &= \mathbf{a}_t + \mathbf{K}_t e_t \\ \mathsf{C}_t &= \mathsf{R}_t - \mathsf{A}_t \mathsf{Q}_t \mathsf{A}_t^T \end{aligned}$$

Having these recursions, we finally define the *filtering density* as

$$\boldsymbol{\theta}_t | \mathcal{D}_t \sim \mathcal{N}\left(\mathbf{m}_t, \mathsf{C}_t\right). \tag{4.13}$$

The KF also allows for the evaluation of the likelihood $p\left(\mathcal{D}_t\right)$. If we take into consideration the prediction step in (4.7) and the formulation of the observation equation as in (4.8), standard normal theory allows us to write the conditional likelihood as

$$\begin{aligned} p\left(y_t | \mathcal{D}_{t-1}\right) &= \mathcal{N}\left(y_t | \mathsf{F}^T \mathbf{a}_t, \mathsf{F}\mathsf{R}_t\mathsf{F}^T + V\right) \\ &= \mathcal{N}\left(y_t | f_t, Q_t\right) \end{aligned}$$

which is the predictive density as detailed in (4.9). Since we now that

$$p\left(\mathcal{D}_t\right) = p\left(y_1\right) \prod_{n=2}^{t} p\left(y_n | \mathcal{D}_{n-1}\right),$$

we can then use the predictive density to calculate $p\left(\mathcal{D}_t\right)$ as

$$p\left(\mathcal{D}_t\right) = p\left(y_1\right) \prod_{n=2}^{t} p\left(y_n | f_n, Q_n\right).$$

A generic algorithm for the KF is presented in Algorithm 4.1.

Example. KF for a $\mathcal{M} = \mathcal{P}\left(1\right)$ NDLM.

Here we consider the state estimation for a realisation of a $\mathcal{M} = \mathcal{P}\left(1\right)$ NDLM with parameters $\boldsymbol{\Phi} = \{W, V\} = \left\{\tau^2, \sigma^2\right\} = \{2, 3\}$ with $N_{obs} = 1000$, and with prior moments $\mathbf{m}_0 = 0$ and $\mathsf{C}_0 = 10$. We can see the result of the state estimation using the KF in Fig-

Algorithm 4.1 Kalman Filter

initialisation Having prior moments $\mathbf{m}_0$ and C_0.

for $t \leftarrow 1$ to N_{obs}

predict set $\mathbf{a}_t = \mathsf{G}\mathbf{m}_{t-1}$ and $\mathsf{R}_t = \mathsf{G}\mathsf{C}_{t-1}\mathsf{G}^T + \mathsf{W}$

update set

$$\begin{aligned} f_t &= \mathsf{F}^T \mathbf{a}_t \\ e_t &= y_t - f_t \\ Q_t &= \mathsf{F}\mathsf{R}_t\mathsf{F}^T + V \\ \mathbf{K}_t &= \mathsf{R}_t\mathsf{F}^T Q_t^{-1} \end{aligned}$$

and calculate moments as

$$\begin{aligned} \mathbf{m}_t &= \mathbf{a}_t + \mathbf{K}_t e_t \\ \mathsf{C}_t &= \mathsf{R}_t - \mathsf{A}_t Q_t \mathsf{A}_t^T \end{aligned}$$

filtering density at time t is

$$\boldsymbol{\theta}_t | \mathcal{D}_t \sim \mathcal{N}(\mathbf{m}_t, \mathsf{C}_t)$$

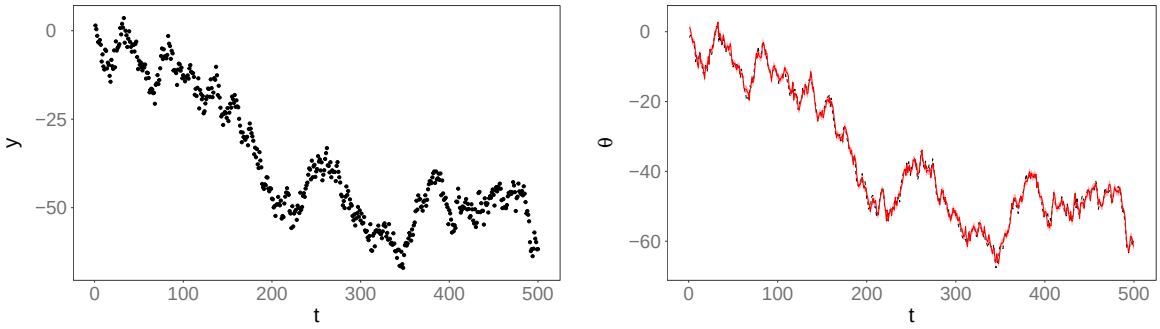
ure 4.1b on the facing page and the observations for the model's realisation in Figure 4.1a on the next page.

To perform a k-step ahead state forecast with the Kalman filter, we can start by defining the moments in the propagation density (4.7) and predictive density (4.9) as a function of the step index k, conditioned on the data up to the forecast starting point t, such that

$$\begin{aligned} \mathbf{a}_t(k) &= \mathrm{E}[\boldsymbol{\theta}_{t+k} | \mathcal{D}_t] \\ \mathsf{R}_t(k) &= \mathrm{Var}[\boldsymbol{\theta}_{t+k} | \mathcal{D}_t] \\ f_t(k) &= \mathrm{E}[y_{t+k} | \mathcal{D}_t] \\ Q_t(k) &= \mathrm{Var}[y_{t+k} | \mathcal{D}_t]. \end{aligned}$$

Assuming that at time t, the starting point of the forecast, we define the moments

$$\begin{aligned} \mathbf{a}_t(0) &= \mathbf{m}_t \\ \mathsf{R}_t(0) &= \mathsf{C}_t, \end{aligned}$$



(a) $y_{1:t}$ (b) $\theta_{0:t}$

Figure 4.1: Observations (*left*), filtering distribution mean (*right, red*) and "true" state (*right, dashed*) for a realisation of $\mathcal{M} = \mathcal{P}(1)$ NDLM with $\sigma^2 = 2.0$ and $\tau^2 = 3.0$ with $N_{obs} = 1000$, and with prior moments $\mathbf{m}_0 = 0$ and $\mathsf{C}_0 = 10$.

we can then recursively calculate the forecast propagation moments as

$$\mathbf{a}_t(k) = \mathsf{G}\mathbf{a}_t(k-1) \tag{4.14}$$

$$\mathsf{R}_t(k) = \mathsf{G}\mathsf{R}_t(k-1)\mathsf{G}^T + \mathsf{W}, \tag{4.15}$$

which will provide the state *forecast distribution* as

$$p(\boldsymbol{\theta}_{t+k}|\mathcal{D}_t) = \mathcal{N}(\mathbf{a}_t(k), \mathsf{R}_t(k)).$$

To perform k-step ahead observation forecasts, we can apply a similar recursive calculation using the predictive density moments, that is

$$f_t(k) = \mathsf{F}^T\mathbf{a}_t(k) \tag{4.16}$$

$$Q_t(k) = \mathsf{F}^T\mathsf{R}_t(k)\mathsf{F} + V. \tag{4.17}$$

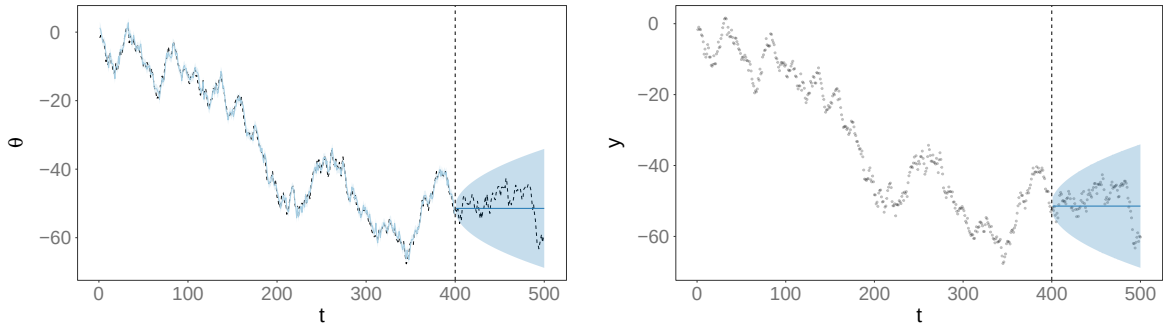
This will allows us to calculate the k-step ahead observation forecast as

$$p(y_{t+k}|\mathcal{D}_t) = \mathcal{N}(f_t(k), Q_t(k)).$$

It is worth noting that the Kalman filter also gives us the filtering, forecast and predictive distribution estimate's uncertainty by providing, respectively, (4.12), (4.15) and (4.17), which can be used to determine, for instance, the estimate's confidence interval.

Example. State and observation forecast in a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM.

Here we consider a realisation of $N_{obs} = 500$ observations of a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with parameter set $\Phi = \{\tau^2, \nu^2\} = \{3, 2\}$. A state and observation k-step ahead ($k = 100$)



(a) $\hat{\theta}_{t+k}$ for $t = 400, k = 100$

(b) $\hat{y}_{t+k}$ for $t = 400, k = 100$

Figure 4.2: Filtering density mean (*left, blue*), state forecast mean (*left, blue line*) and forecast 90% CI (*left, shaded blue*). "True" state as black dashed line (*left*). Observations (*right*) and k-step ahead observation forecasts (*right, blue line*) and forecast 90% CI (*right, shaded blue*) for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM using the KF

forecast was performed at time $t = 400$, using the recursions specified in (4.14), (4.15), (4.16) and (4.17). The state forecast can be seen in Figure 4.2a and the observation forecast can be viewed in Figure 4.2b.

4.1 Using Singular Value Decomposition

In Wang *et al.* (1992) an algorithm is presented in which the Kalman filter recursions can be calculated using Singular Value Decomposition (SVD). Although theoretically impossible, in practice, when calculating the updated second moment (4.12) using the standard KF recursions, a computationally non-positive definite value may arise. According to Wang *et al.* (1992) this method can improve the numerical stability of the estimations.

The notation for SVD used will that considering a $m \times n$ (with $m \geq n$) matrix A the SVD is a factorisation of three matrices (Wang *et al.* (1992)) such that

$$\mathsf{A} = \mathsf{U}\Lambda\mathsf{V}^T, \qquad \Lambda = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}$$

where $\mathsf{S} = \text{diag}(\sigma_1, \ldots, \sigma_r)$ and U is a $m \times m$ orthogonal matrix and V a $n \times n$ matrix. According to Wang *et al.* (1992), a special case occurs when A is positive definite, in which we have

$$\mathsf{A} = \mathsf{U} \begin{bmatrix} \mathsf{S} \\ 0 \end{bmatrix} \mathsf{V}^T$$

with S an $n \times n$ diagonal matrix. Another special case is noted in Wang *et al.* (1992), which occurs when A is symmetric positive definite, in which case we have

$$\mathsf{A} = \mathsf{USU}^T = \mathsf{U}D^2\mathsf{U}^T.$$

. If, at time t, we assume that we have the SVD decomposition of $\mathsf{C}_{t|t}$, that is

$$\mathsf{C}_{t|t} = \mathsf{U}_{t|t}D^2_{t|t}\mathsf{U}^T_{t|t},$$

we can then write (4.6) as

$$\begin{aligned} \mathsf{C}_{t+1|t} &= \mathsf{G}\mathsf{C}_{t|t}\mathsf{G}^T + \mathsf{W} \\ &= \mathsf{G}\mathsf{U}_{t-1|t-1}D^2_{t-1|t-1}\mathsf{U}^T_{t-1|t-1}\mathsf{G}^T + \mathsf{W}. \end{aligned}$$

The objective is then to find an SVD decomposition with components $\mathsf{U}_{t+1|t}, D^2_{t+1|t}$ such that

$$\mathsf{C}_{t+1|t} = \mathsf{U}_{t+1|t}D^2_{t+1|t}\mathsf{U}^T_{t+1|t}.$$

Wang *et al.* (1992) defines the following matrix

$$\begin{bmatrix} D_{t|t}\mathsf{U}^T_{t|t}\mathsf{G}^T \\ \sqrt{\mathsf{W}^T} \end{bmatrix}, \tag{4.18}$$

and by calculating the SVD of (4.18) the result is

$$\begin{bmatrix} D_{t|t}\mathsf{U}^T_{t|t}\mathsf{G}^T \\ \sqrt{\mathsf{W}^T} \end{bmatrix} = \mathsf{U}'_t \begin{bmatrix} D'_t \\ 0 \end{bmatrix} \left(\mathsf{V}'_t\right)^T.$$

Wang *et al.* (1992) continues by multiplying each side by its transpose, reformulating into

$$\begin{aligned} \begin{bmatrix} D_{t|t}\mathsf{U}^T_{t|t}\mathsf{G}^T \\ \sqrt{\mathsf{W}^T} \end{bmatrix}^T \begin{bmatrix} D_{t|t}\mathsf{U}^T_{t|t}\mathsf{G}^T \\ \sqrt{\mathsf{W}^T} \end{bmatrix} &= \left\{ \mathsf{U}'_t \begin{bmatrix} D'_t \\ 0 \end{bmatrix} \left(\mathsf{V}'_t\right)^T \right\}^T \mathsf{U}'_t \begin{bmatrix} D'_t \\ 0 \end{bmatrix} \left(\mathsf{V}'_t\right)^T \\ \mathsf{G}\mathsf{U}_{t|t}D^2_{t|t}\mathsf{U}^T_{t|t}\mathsf{G}^T + \underbrace{\sqrt{\mathsf{W}}\left(\sqrt{\mathsf{W}}\right)^T}_{\mathsf{W}} &= \mathsf{V}'_t \begin{bmatrix} D'_t & 0 \end{bmatrix} \underbrace{\left(\mathsf{U}'_t\right)^T \mathsf{U}'_t}_{I} \begin{bmatrix} D'_t \\ 0 \end{bmatrix} \left(\mathsf{V}'_t\right)^T \\ \underbrace{\mathsf{G}\mathsf{U}_{t|t}D^2_{t|t}\mathsf{U}^T_{t|t}\mathsf{G}^T + \mathsf{W}}_{\mathsf{C}_{t+1|t}} &= \underbrace{\mathsf{V}'_t}_{\mathsf{U}_{t+1|t}} \underbrace{\left(D'_t\right)^2}_{D^2_{t+1|t}} \underbrace{\left(\mathsf{V}'_t\right)^T}_{\mathsf{U}^T_{t+1|t}} \end{aligned}$$

From the above definition of $\mathsf{C}_{t+1|t}$, Wang *et al.* (1992) notes that the required components for the SVD are

$$\begin{aligned} \mathsf{U}_{t+1|t} &= \mathsf{V}'_t \\ D_{t+1|t} &= D'_t \end{aligned}$$

Regarding the *update* steps of the KF, to use SVD, the above decompositions of $\mathsf{C}_{t+1|t+1}$ and $\mathsf{C}_{t+1|t}$ are applied:

$$\mathsf{C}^{-1}_{t+1|t+1} = \left(\mathsf{U}_{t+1|t+1} D^2_{t+1|t+1} \mathsf{U}^T_{t+1|t+1}\right)^{-1}$$

Since

$$\mathsf{C}^{-1}_{t+1|t+1} = \mathsf{C}^{-1}_{t+1|t} + \mathsf{F}^T \mathsf{R}^{-1} \mathsf{F},$$

it follows that

$$\begin{aligned} \mathsf{C}^{-1}_{t+1|t+1} &= \underbrace{\left(\mathsf{U}_{t+1|t} D^2_{t+1|t} \mathsf{U}^T_{t+1|t}\right)^{-1}}_{\mathsf{C}^{-1}_{t+1|t}} + \mathsf{F}^T \mathsf{R}^{-1}_{t+1} \mathsf{F} \\ &= \left(\mathsf{U}^T_{t+1|t}\right)^{-1} D^{-2}_{t+1|t} \mathsf{U}^{-1}_{t+1|t} + \left(\mathsf{U}^T_{t+1|t}\right)^{-1} \mathsf{U}^T_{t+1|t} \mathsf{F}^T \mathsf{R}^{-1}_{t+1} \mathsf{F} \mathsf{U}_{t+1|t} \mathsf{U}^{-1}_{t+1|t} \\ &= \mathsf{U}^{-T}_{t+1|t} \left(D^{-2}_{t+1|t} + \mathsf{U}^T_{t+1|t} \mathsf{F}^T \mathsf{R}^{-1}_{t+1} \mathsf{F} \mathsf{U}_{t+1|t}\right) \mathsf{U}^{-1}_{t+1|t}. \end{aligned}$$

If we consider the Cholesky decomposition of the inverse of the covariance matrix V as $\mathsf{R}^{-1}_{t+1} = \mathsf{L}_{t+1} \mathsf{L}^T_{t+1}$, in the univariate observation case, which is the one we are considering throughout, we can simply define $L_{t+1} = \sqrt{1/V}$. Wang *et al.* (1992) continues by creating the matrix

$$\begin{bmatrix} \mathsf{L}^T_{t+1} \mathsf{F} \mathsf{U}_{t+1|t} \\ D^{-1}_{t+1|t} \end{bmatrix}$$

and calculating the SVD of the above, resulting in

$$\begin{bmatrix} \mathsf{L}^T_{t+1} \mathsf{F} \mathsf{U}_{t+1|t} \\ D^{-1}_{t+1|t} \end{bmatrix} = \bar{\mathsf{U}}'_{t+1} \begin{bmatrix} \bar{D}'_{t+1} \\ 0 \end{bmatrix} \left(\bar{\mathsf{V}}'_{t+1}\right)^T$$

from which

$$\mathsf{C}^{-1}_{t+1|t+1} = \left(\mathsf{U}^T_{t+1|t}\right)^{-1} \bar{\mathsf{V}}'_{t+1} \left(\bar{D}'_{t+1}\right)^2 \left(\bar{\mathsf{V}}'_{t+1}\right)^T \mathsf{U}^{-1}_{t+1|t}.$$

The decomposition can then be written as

$$\begin{aligned} \mathsf{U}_{t+1|t+1} &= \mathsf{U}_{t+1|t}\bar{\mathsf{V}}'_{t+1} = \mathsf{V}'_t\bar{\mathsf{V}}'_{t+1} \\ D_{t+1|t+1} &= \left(\bar{D}'_{t+1}\right).^{-1} \end{aligned}$$

As for the Kalman gain, Wang *et al.* (1992) calculates it according to

$$\begin{aligned} \mathbf{K}_{t+1} &= \mathsf{C}_{t+1|t+1}\mathsf{F}^T\mathsf{R}_{t+1}^{-1} \\ &= \mathsf{U}_{t+1|t+1}D^2_{t+1|t+1}\mathsf{U}^T_{t+1|t+1}\mathsf{F}^T\mathsf{R}_{t+1}^{-1}. \end{aligned}$$

A general algorithm for Wang *et al.* (1992) formulation of the KF-SVD is presented in Algorithm 4.2.

Example. Kalman Filter SVD for a NDLM

Consider a realisation of a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(250, 2)\}$ NDLM, that is, a locally constant component $W = 1.2$ and a seasonal component with 2 harmonics of period $p = 250$ and parameters

$$W = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.7 \end{bmatrix}, \qquad V = 1.5,$$

with $N_{obs} = 1000$ observations. The data is presented in Figure 4.3a on page 41 and the state estimation for each state component are presented in Figures 4.3b, 4.3c and 4.3d. In this particular example, since we are dealing with a simple model with simulated observations, we wouldn't expect the standard KF to incur numerical instability. The Mean Squared Error (MSE), calculated as

$$MSE = \frac{1}{N_{obs}}\sum_{t=1}^{N_{obs}}\left(\mathbf{m}_t - \boldsymbol{\theta}_t\right)^2,$$

where $\boldsymbol{\theta}_t$ is "true" value from the realisation, is identical[1] for all state components as we can see from table 4.1. The main difference is however the computational cost, which is higher for the SVD implementation. This is expected due to additional steps necessary regarding SVD decomposition as described in this section.

[1] In theory, it is expected for the standard and SVD KF to produce the same estimation results. In practice, due to numerical instability, this might not be the case.

Algorithm 4.2 SVD based Kalman Filter (KF-SVD)

initialise with $\mathbf{m}_0, \mathsf{C}_0$. Assuming C_0 diagonal, $\mathsf{U}_0 = \mathbf{I}$ and $\mathsf{D}_0 = \mathsf{C}_0$.

for $t \leftarrow 1$ to N_{obs}

calculate (assuming $L_t = \sqrt{1/V}$) V_t' and D_t' from the SVD decomposition of

$$\begin{bmatrix} L_t^T \mathsf{F} \mathsf{U}_t \\ D_t^{-1} \end{bmatrix}$$

update $\mathsf{U}_{t+1|t}$ and $D_{t+1|t}$ according to

$$\begin{aligned} \mathsf{U}_{t+1|t} &= \mathsf{U}_t \mathsf{V}_t' \\ D_{t+1|t} &= \left(D_t'\right)^{-1} \end{aligned}$$

calculate Kalman gain

$$\mathbf{K}_t = \mathsf{U}_{t+1|t} D_{t+1|t}^2 \mathsf{U}_{t+1|t}^T \mathsf{F}^T \mathsf{L}_t \mathsf{L}_t^T$$

predict step

$$\hat{\boldsymbol{\theta}}_{t+1|t} = \hat{\boldsymbol{\theta}}_t + \mathbf{K}_t \left(y_t - \mathsf{F}\boldsymbol{\theta}_t\right)$$

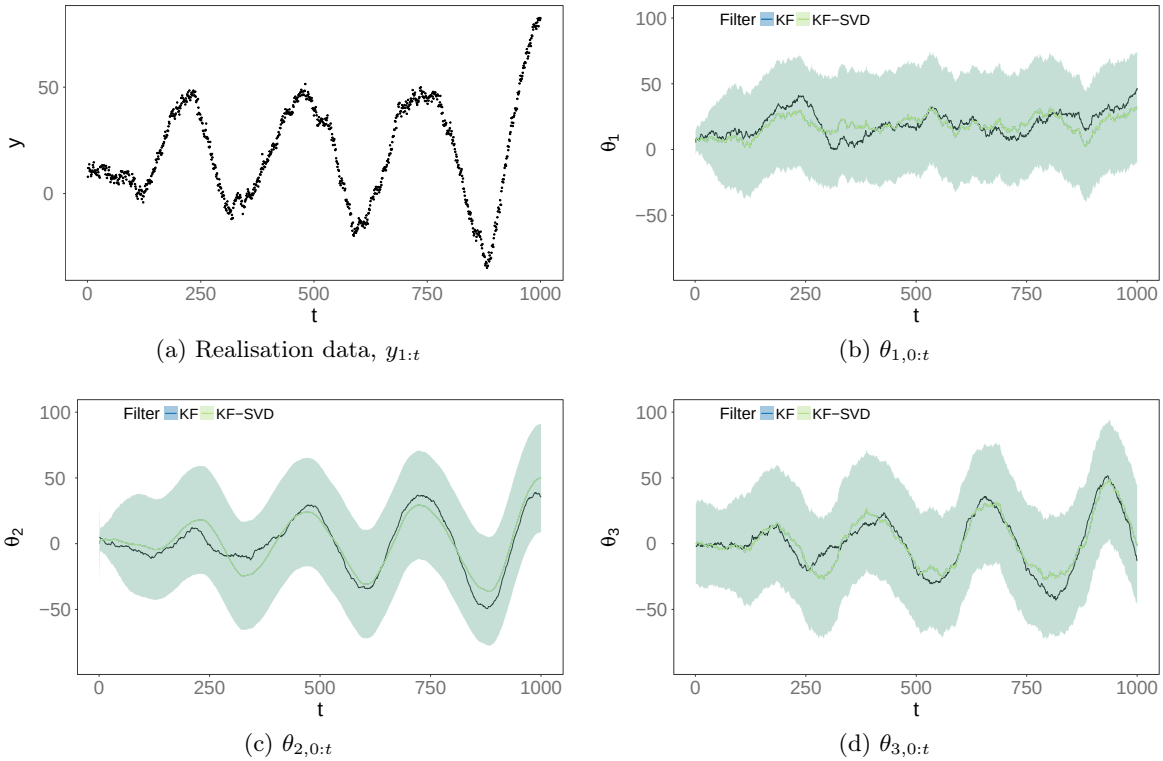
correct step. Calculate V_t' and D_t' from the SVD decomposition of

$$\begin{bmatrix} D_{t+1|t} \mathsf{U}_{t+1|t}^T \mathsf{G}^T \\ \sqrt{\mathsf{W}^T} \end{bmatrix}$$

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{t+1} &= \mathsf{G} \hat{\boldsymbol{\theta}}_{t+1|t} \\ \mathsf{U}_{t+1} &= \mathsf{V}_t' \\ D_{t+1} &= D_t' \end{aligned}$$

Filter	**MSE**			**time** (ms)
	θ_1	θ_2	θ_3	
KF	54.455	53.415	47.374	10.0
KF-SVD	54.455	53.415	47.374	17.0

Table 4.1: Summary of MSE and computation times for KF and KF-SVD for a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(250, 2)\}$ Gaussian DLM with $N_{obs} = 1000$.



(a) Realisation data, $y_{1:t}$

(b) $\theta_{1,0:t}$

(c) $\theta_{2,0:t}$

(d) $\theta_{3,0:t}$

Figure 4.3: Realisation of a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(250, 2)\}$ Gaussian DLM and filtering density estimation using standard and SVD Kalman filters. Solid colour lines represent filtering density mean and shaded areas 90% CI. Solid black line represents the "true" (realisation) state.

4.2 Extended Kalman Filter

In the case where the model is non-linear, the Kalman filter recursions examined previously in this chapter cannot be applied as they depend on the joint normality of state and observations. If we recall that a general state-space model can be formulated as in (2.1) and (2.2), that is

$$\begin{aligned} y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi}_t &\sim f(y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi}_t) \\ \boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi}_t &\sim g(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi}_t), \end{aligned}$$

a method proposed by Jazwinski (1966); Maybeck (1979) consists of approximating the non-linear functions by means of a first order Taylor series expansion around the estimates at time t and subsequently applying the KF recursion as in the NDLM case. This method, the Extended Kalman Filter (EKF) will then approximate the posterior $p(\boldsymbol{\theta}_t|\mathcal{D}_t)$

by approximating the model to be the linearised form

$$\begin{aligned} y_t &= f^*(\boldsymbol{\theta}_t, t) + \tilde{v}_t, \qquad \tilde{v}_t \sim \mathcal{N}\left(0, \tilde{V}_t\right) \\ \boldsymbol{\theta}_t &= g^*(\boldsymbol{\theta}_{t-1}, t-1) + \tilde{w}_t, \qquad \tilde{w}_t \sim \mathcal{N}\left(0, \tilde{\mathsf{W}}_t\right), \end{aligned}$$

where $f^*(\cdot)$ and $g^*(\cdot)$ are differentiable but not required to be linear functions of the state, with a filtering density of

$$p(\boldsymbol{\theta}_t|\mathcal{D}_t) \approx \mathcal{N}(\mathbf{m}_t, \mathsf{C}_t).$$

Additionally, the following relations can be defined (Arulampalam *et al.* (2002)), in some way drawing parallels with the KF:

$$\begin{aligned} \mathbf{a}_t &= g^*(\mathbf{m}_{t-1}) \\ \mathsf{R}_t &= \hat{\mathsf{G}}_t \mathsf{C}_{t-1} \hat{\mathsf{G}}^T + \tilde{\mathsf{W}}_{t-1} \\ \mathbf{m}_t &= \mathbf{a}_t + K_t(y_t - f^*(\mathbf{a}_t)) \\ \mathsf{C}_t &= \mathsf{R}_t - K_t \hat{\mathsf{F}}^T \mathsf{R}_t, \end{aligned}$$

where $\hat{\mathsf{F}}_t$ and $\hat{\mathsf{G}}_t$ represent the state-space linearised functions

$$\begin{aligned} \hat{\mathsf{F}}_t &= \left.\frac{df^*(\boldsymbol{\theta})}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\mathbf{m}_{t-1}} \\ \hat{\mathsf{G}}_t &= \left.\frac{dg^*(\boldsymbol{\theta})}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\mathbf{m}_{t-1}}, \end{aligned}$$

and where

$$\begin{aligned} Q_t &= \hat{\mathsf{H}}_t \mathsf{R}_t \hat{\mathsf{H}}_t^T + \tilde{V}_t \\ K_t &= \mathsf{R}_t \hat{\mathsf{H}}_t^T Q_t^{-1}, \end{aligned}$$

with $\tilde{V}_t$ and $\tilde{\mathsf{W}}_t$ representing the variances of the additive noise for the linearised model.

We have seen, however, in Section 2.1, that the DGLM is a specific instance of the more general state-space model, where

$$g(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi}_t) = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathsf{W})$$

and $f(y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi}_t)$ is a distribution from the exponential family in the form

$$y_t \sim p(y_t|\eta_t) = \exp\left\{\frac{z(y_t)\,\eta_t - b(\eta_t)}{a(\phi_t)} + c(y_t, \phi_t)\right\}.$$

In this section we will use a formulation presented in Gamerman (1998); Fahrmeir (1992), which aims specifically at approximating DGLMs to a linearised form using a generalisation of the EKF. Gamerman (1998); Fahrmeir (1992) define an *adjusted model* which will act as a linear approximation to the NDLM. Considering the DGLM mean as defined in (2.6) as

$$\mu_t = b'(\eta_t),$$

and considering, as in (2.9), that

$$g(\mu_t) = h^{-1}(\mu_t) = \lambda_t = \eta_t$$

where η_t is the natural parameter.

Since $g(\cdot)$ and $b(\cdot)$ are known from Section 2.3.2, the adjusted model can be easily calculated for each of the non-linear DGLM instances. Using this model, and considering that for the DGLM the underlying state will have a linear evolution, we can write, as in the KF, the filtering density at time $t-1$ as

$$\begin{aligned} p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) &\approx \mathcal{N}\left(\mathsf{G}\mathbf{m}_{t-1}, \mathsf{G}\mathsf{C}_{t-1}\mathsf{G}^T + \mathsf{W}\right) \\ &= \mathcal{N}(\mathbf{a}_t, \mathsf{R}_t). \end{aligned}$$

And define the linearised $\tilde{\lambda}_t$ as

$$\lambda_t = \mathsf{F}^T\mathbf{a}_t, \tag{4.19}$$

Gamerman (1998) formulates the *adjusted model* as

$$\begin{aligned} y_t \approx \tilde{y}_t &= \lambda_t + (y_t - \mu_t)\, g'(\mu_t) + \tilde{\nu}_t \\ \tilde{\nu}_t &\sim \mathcal{N}\left(0, \tilde{V}_t\right) \\ \tilde{V}_t &= b''(\eta_t)\{g'(\mu_t)\}^2. \end{aligned}$$

With these quantities we can then calculate the predictive density of the adjusted model as

$$\begin{aligned} \tilde{y}_t|\mathcal{D}_{t-1} &\sim \mathcal{N}(f_t, Q_t) \\ f_t &= \mathsf{F}^T\mathbf{a}_t \\ Q_t &= \mathsf{F}^T\mathsf{R}_t\mathsf{F} + \tilde{V}_t. \end{aligned}$$

Similarly to the KF, as detailed in (4.11) and (4.12), the moment update will be defined

by

$$\begin{aligned}\mathbf{m}_t &= \mathbf{a}_t + A_t\left(\tilde{y}_t - f_t\right)\\ \mathsf{C}_t &= \mathsf{R}_t - \mathbf{A}_t Q_t \mathbf{A}_t^T\\ \mathbf{A}_t &= \mathsf{R}_t F.\end{aligned}$$

This method, while aiming at approximating the filtering density of the linearised $p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right)$, leads to poor results in the presence of strong non-linearities due to mismatch between the mode of the true filtering density and linearised filtering density mode. In the case of filters such as the Iterated Extended Kalman Filter (IEKF, Jazwinski (1970)) this problem can be mitigated by using an iterative Newton-Raphson method until the approximated density converges. In the case of the adjusted model presented here, as mentioned in Gamerman (1998); Fahrmeir (1992), the approximation can be improved by, in an offline fashion (Fahrmeir, 1992), iteratively applying a Kalman smoother to calculate the smoothed moments $\tilde{\mathbf{m}}_t, \tilde{\mathsf{C}}_t$ until mode convergence is achieved. This value could then be used to calculate (4.19).

A general algorithm for the adjusted model EKF (without the iterative mode search) is presented in Algorithm 4.3.

Example. Linearised Poisson DLM

If we consider a Poisson DLM in the form presented in Section 2.3.2, that is, with observational model

$$\begin{aligned}y_t|\eta_t &\sim \mathrm{Po}\left(e^{\eta_t}\right)\\ \eta_t &= \mathsf{F}^T\boldsymbol{\theta}_t,\end{aligned}$$

and canonical quantities

$$\begin{aligned}z\left(y_t\right) &= y_t\\ a\left(\phi_t\right) &= 1\\ b\left(\eta_t\right) &= e^{\eta_t}\\ c\left(y_t, \phi_t\right) &= \log\left(\frac{1}{y_t!}\right),\end{aligned}$$

Algorithm 4.3 Adjusted Model Extended Kalman Filter (EKF)

initialisation

set state prior moments $\mathbf{m}_0$ and C_0

for $t \leftarrow 1$ to N_{obs}

calculate

$$\begin{aligned}\mathbf{a}_t &= \mathsf{G}\mathbf{m}_{t-1}\\ \mathsf{R}_t &= \mathsf{G}\mathsf{C}_{t-1}\mathsf{G}^T + \mathsf{W}\\ \tilde{\lambda}_t &= \mathsf{F}^T\mathbf{a}_t\end{aligned}$$

calculate adjusted observation model according to the DGLM class

$$\begin{aligned}\tilde{y}_t &= \lambda_t + (y_t - \mu_t)\, g'(\mu_t)\\ \tilde{V}_t &= b''(\eta_t)\{g'(\mu_t)\}^2\end{aligned}$$

calculate adjusted model predictive moments

$$\begin{aligned}f_t &= \mathsf{F}^T\mathbf{a}_t\\ Q_t &= \mathsf{F}^T\mathsf{R}_t\mathsf{F}_t + \tilde{V}_t\end{aligned}$$

calculate updated state posterior moments

$$\begin{aligned}\mathbf{m}_t &= \mathbf{a}_t + \mathbf{A}_t(\tilde{y}_t - f_t)\\ \mathsf{C}_t &= \mathsf{R}_t - \mathbf{A}_t Q_t \mathbf{A}_t^T\end{aligned}$$

with

$$\mathbf{A}_t = \mathsf{R}_t\mathsf{F}$$

to define the adjusted model we use the quantities calculated in (2.20) and (2.21):

$$\begin{aligned}b'(\eta_t) &= e^{\eta_t}\\ b''(\eta_t) &= e^{\eta_t}\\ \mu_t &= b'(\eta_t) = e^{\eta_t}.\end{aligned}$$

Since

$$\lambda_t = \mathsf{F}^T\boldsymbol{\theta}_t,$$

the adjusted observational model will take the form

$$\begin{aligned} y_t \approx \tilde{y}_t &= \lambda_t + (y_t - \mu_t)\, g'(\mu_t) + \tilde{\nu}_t \\ &= \lambda_t + (y_t - \lambda_t)\, \lambda_t^{-1} + \tilde{\nu}_t, \end{aligned}$$

with

$$\begin{aligned} \tilde{\nu}_t &\sim \mathcal{N}\left(0, \tilde{V}_t\right) \\ \tilde{V}_t &= b''(\eta_t)\, \{g'(\mu_t)\}^2 \\ &= \lambda_t \lambda_t^{-2} = \lambda_t^{-1}. \end{aligned}$$

Consequently, the adjusted PoDLM will take the form

$$\begin{aligned} p(y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi}) &\approx \mathcal{N}\left(\lambda_t + (y_t - \lambda_t)\, \lambda_t^{-1}, \lambda_t^{-1}\right) \\ \lambda_t &= \mathsf{F}^T \boldsymbol{\theta}_t \\ p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi}) &= \mathcal{N}(\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}). \end{aligned}$$

Example. Linearised Binomial Model

We now consider a Binomial DLM in the form of Section 2.3.3, that is, with observational model

$$\begin{aligned} y_t|\eta_t &\sim \text{Binom}\{\text{logit}^{-1}(\eta_t\}, n_t) \\ \eta_t &= \mathsf{F}^T \boldsymbol{\theta}_t \end{aligned}$$

with the canonical quantities

$$\begin{aligned} z(y_t) &= \frac{y_t}{n_t} \\ \eta_t &= \log\left(\frac{p_t}{1-p_t}\right) \\ a(\phi_t) &= \phi_t^{-1} = n_t^{-1} \\ b(\eta_t) &= \log(1 + e^{\eta_t}) \\ c(y_t, \phi_t) &= \log \binom{n_t}{y_t}. \end{aligned}$$

The adjusted observational model can be defined by the quantities, previously calcu-

lated in (2.23) and (2.24):

$$\begin{aligned} b'(\eta_t) &= n_t \text{logit}^{-1}\{\eta_t\} \\ &= n_t p_t \\ b''(\eta_t) &= n_t p_t (1-p_t) \\ \mu_t &= b'(\eta_t) = n_t p_t. \end{aligned}$$

The adjusted observational model can then be expressed as

$$\begin{aligned} y_t \approx \tilde{y}_t &= \lambda_t + (y_t - \mu_t) g'(\mu_t) + \tilde{\nu}_t \\ &= y_t n_t^{-1} \exp\{-\lambda_t\} (1+\exp\{\lambda_t\})^2 - \exp\{\lambda_t\} - 1 + \lambda_t + \tilde{\nu}_t, \end{aligned}$$

with

$$\begin{aligned} \tilde{\nu}_t &\sim \mathcal{N}\left(0, \tilde{V}_t\right) \\ \tilde{V}_t &= b''(\eta_t)\{g'(\mu_t)\}^2 \\ &= n_t^{-1} \exp\{-\lambda_t\} (1+\exp\{\lambda_t\})^2. \end{aligned}$$

Consequently, the adjusted BDLM will take the form

$$\begin{aligned} p(y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi}) &\approx \mathcal{N}\left(y_t n_t^{-1} \exp\{-\lambda_t\} (1+\exp\{\lambda_t\})^2 - \exp\{\lambda_t\} - 1 + \lambda_t, \tilde{V}_t\right) \\ \tilde{V}_t &= n_t^{-1} \exp\{-\lambda_t\} (1+\exp\{\lambda_t\})^2 \\ \lambda_t &= \mathsf{F}^T \boldsymbol{\theta}_t \\ p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi}) &= \mathcal{N}(\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}). \end{aligned}$$

Chapter 5

Conjugate Filtering

5.1 Conjugate filtering

As we've seen in Chapter 4, for the NDLM we can use an exact solution for the filtering problem. However, in all other non-linear DGLMs an analytic solution will not be available. A method proposed in West & Harrison (1997), called *conjugate filtering*, assumes partially specified distributions and an approximation using linear Bayes to achieve a recursive approximation to the state's posterior given the observations, $\boldsymbol{\theta}_t|\mathcal{D}_t$ using conjugate updates as defined in Section 5.2.

We start by considering a DGLM in the canonical form of (2.5), that is

$$\begin{aligned} y_t \sim p\left(y_t|\eta_t\right) &= \exp\left\{\frac{z\left(y_t\right)\eta_t - b\left(\eta_t\right)}{a\left(\phi_t\right)} + c\left(y_t,\phi_t\right)\right\} \\ \eta_t &= g^{-1}\left(\lambda_t\right) \\ \lambda_t &= \mathsf{F}^T\boldsymbol{\theta}_t \\ \boldsymbol{\theta}_t &\sim \mathcal{N}\left(\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}\right) \\ &= \mathsf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim \mathcal{N}\left(\mathbf{0}, \mathsf{W}\right), \end{aligned}$$

with a state prior of $\boldsymbol{\theta}_0|\mathcal{D}_0 \sim [\mathbf{m}_0, \mathsf{C}_0]$. In West & Harrison (1997), the notation used indicates μ_t as the *observation mean*, related to the natural parameter η_t as described in section (2.3) via

$$\begin{aligned} \mu_t &= b'\left(\eta_t\right) \\ \Sigma &= a\left(\phi_t\right) b''\left(\eta_t\right). \end{aligned}$$

Additionally, η_t relates to the linear function $\lambda_t = \mathsf{F}^T\boldsymbol{\theta}_t$ via $\lambda_t = g\left(\eta_t\right)$. West & Harrison

(1997) stipulates that since $b'(\cdot)$ and $g(\cdot)$ are bijective we can work interchangeably with μ_t, η_t or λ_t.

In DGLMs, in general, we cannot assume joint normality between the state and observations as we did in the KF derivations. The observation equation is generally non-linear (Poisson, Binomial, *etc.*) and the observation mean μ_t might be a non-linear function of the state, and as mentioned previously, an exact analytic solution to the filtering problem will not be available and an approximation must be calculated. Here, we will use West & Harrison (1997) notation, where

$$[\boldsymbol{\mu}, \Sigma]$$

indicates a mean, $\boldsymbol{\mu}$, and a variance, Σ on a otherwise unspecified distribution.

West & Harrison (1997) starts by assuming that at time $t-1$ we have the state posterior $\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1} \sim [\mathbf{m}_{t-1}, \mathsf{C}_{t-1}]$ such as in the NDLM, and that (as specified in Chapter 2.1) we have independence between $\boldsymbol{\theta}_t$ and ω_t so that we can then write the state's partial prior in the form (without assuming normality) of (4.7). That is, from (4.5) and (4.6) we can write

$$\boldsymbol{\theta}_t|\mathcal{D}_{t-1} \sim \left[\underbrace{\mathsf{G}\mathbf{m}_{t-1}}_{\mathbf{a}_t}, \underbrace{\mathsf{G}\mathsf{C}_{t-1}\mathsf{G}^T + \mathsf{W}}_{\mathsf{R}_t}\right]. \tag{5.1}$$

Since $\lambda_t = \mathsf{F}^T\boldsymbol{\theta}_t$ is a linear function of $\boldsymbol{\theta}_t$ and assuming the moments in (5.1), we obtain

$$\begin{aligned} \mathrm{E}[\lambda_t|\mathcal{D}_{t-1}] &= f_t = \mathsf{F}^T\mathbf{a}_t \\ \mathrm{Var}[\lambda_t|\mathcal{D}_{t-1}] &= q_t = \mathsf{F}^T\mathsf{R}_t, \end{aligned}$$

and the covariance between $\boldsymbol{\theta}_t$ and λ_t is

$$\mathrm{Cov}[\lambda_t, \boldsymbol{\theta}_t|\mathcal{D}_{t-1}] = \mathsf{F}^T\mathsf{R}_t.$$

Hence, we have the following (*partially* specified) joint prior

$$\left(\begin{array}{c|c} \boldsymbol{\theta}_t & \\ \lambda_t & \mathcal{D}_{t-1} \end{array}\right) \sim \left[\begin{pmatrix} \mathbf{a}_t \\ f_t \end{pmatrix}, \begin{bmatrix} \mathsf{R}_t & \mathsf{R}_t\mathsf{F} \\ \mathsf{F}^T\mathsf{R}_t & q_t \end{bmatrix}\right], \tag{5.2}$$

which strikes some similarities with the Kalman filter derivation in (4.10).

To determine the (partially specified) observation one-step ahead forecast, since the observation model depends on the state $\boldsymbol{\theta}_t$ through the natural parameter, $\eta_t = g^{-1}(\lambda_t)$, it can be specified via the natural parameter's prior $\eta_t|\mathcal{D}_{t-1}$. As West & Harrison (1997)

notes, this is specified in terms of the moments of $\lambda_t = g(\eta_t)$ as indicated in (5.2):

$$\lambda_t | \mathcal{D}_{t-1} \sim [f_t, q_t] .$$

West & Harrison (1997) states that since the prior is partially specified, we can choose any prior distribution of η_t based on the specified mean and variance. Based on Section 5.2 we choose the conjugate prior[1] for η_t as specified in (5.11)

$$p(\eta_t | \mathcal{D}_{t-1}) = CP(r_t, s_t)$$

with the conditions

$$\begin{aligned} \mathrm{E}[g(\eta_t) | \mathcal{D}_{t-1}] &= f_t \\ \mathrm{Var}[g(\eta_t) | \mathcal{D}_{t-1}] &= q_t. \end{aligned}$$

Assuming the observation's one-step ahead forecast expression in (5.13) we can then use the observation y_t to calculate the posterior expression for η_t as in (5.14). Given that we know the mean and variance of λ_t are

$$\lambda_t | \mathcal{D}_t \sim \left[\underbrace{\mathrm{E}[g(\eta_t) | \mathcal{D}_t]}_{f_t^\star}, \underbrace{\mathrm{Var}[g(\eta_t) | \mathcal{D}_t]}_{q_t^\star} \right] \tag{5.3}$$

West & Harrison (1997) notes that because the posterior $\boldsymbol{\theta}_t | \mathcal{D}_t$ is only partially specified, we only need the mean and variance to continue the approximation:

$$\mathbf{m}_t = \mathrm{E}[\boldsymbol{\theta}_t | \mathcal{D}_t] = \mathrm{E}[\mathrm{E}[\boldsymbol{\theta}_t | \lambda_t, \mathcal{D}_{t-1}] | \mathcal{D}_t] \tag{5.4}$$

$$\mathsf{C}_t = \mathrm{Var}[\boldsymbol{\theta}_t | \mathcal{D}_t] = \mathrm{Var}[\mathrm{E}[\boldsymbol{\theta}_t | \lambda_t, \mathcal{D}_{t-1}] | \mathcal{D}_t] + \mathrm{E}[\mathrm{Var}[\boldsymbol{\theta}_t | \lambda_t, \mathcal{D}_{t-1}] | \mathcal{D}_t] . \tag{5.5}$$

It is clear from the above we need the mean and variance of $\boldsymbol{\theta}_t | \lambda_t, \mathcal{D}_{t-1}$. As we've seen previously, this joint distribution is only partially specified so we are not able to calculate them analytically, however, the method described in West & Harrison (1997) applies linear Bayes (with regard to a quadratic loss function for moment estimation) to calculate the

[1] West & Harrison (1997) uses f_t as the mode and q_t as the curvature for the conjugate prior.

optimal estimate of $\mathrm{E}\left[\boldsymbol{\theta}_t|\lambda_t, \mathcal{D}_{t-1}\right]$ and $\mathrm{Var}\left[\boldsymbol{\theta}_t|\lambda_t, \mathcal{D}_{t-1}\right]$ we have[2] the final result:

$$\widehat{\mathrm{E}}\left[\boldsymbol{\theta}_t|, \lambda_t, \mathcal{D}_{t-1}\right] = \mathbf{a}_t + \mathsf{R}_t\mathsf{F}\frac{(\lambda_t - f_t)}{q_t} \tag{5.6}$$

$$\widehat{\mathrm{Var}}\left[\boldsymbol{\theta}_t|, \lambda_t, \mathcal{D}_{t-1}\right] = \mathsf{R}_t - \frac{\mathsf{R}_t\mathsf{F}\mathsf{F}^T\mathsf{R}_t}{q_t}, \tag{5.7}$$

that is,

$$\boldsymbol{\theta}_t|\lambda_t, \mathcal{D}_{t-1} \sim \left[\mathbf{a}_t + \mathsf{R}_t\mathsf{F}\frac{(\lambda_t - f_t)}{q_t}, \mathsf{R}_t - \frac{\mathsf{R}_t\mathsf{F}\mathsf{F}^T\mathsf{R}_t}{q_t}\right].$$

This however, will give an approximation to the moments. Combining (5.4), (5.5), (5.6) and (5.7) we have:

$$\begin{aligned}
\mathbf{m}_t &= \mathrm{E}\left[\widehat{\mathrm{E}}\left[\boldsymbol{\theta}_t|, \lambda_t, \mathcal{D}_{t-1}\right]|\mathcal{D}_t\right] \\
&= \mathrm{E}\left[\mathbf{a}_t + \mathsf{R}_t\mathsf{F}\frac{(\lambda_t - f_t)}{q_t}|\mathcal{D}_t\right] \\
&= \mathbf{a}_t + \mathsf{R}_t\mathsf{F}\frac{1}{q_t}\left(\mathrm{E}\left[\lambda_t|\mathcal{D}_t\right] - f_t\right) \\
\mathsf{C}_t &= \mathrm{Var}\left[\widehat{\mathrm{E}}\left[\boldsymbol{\theta}_t|, \lambda_t, \mathcal{D}_{t-1}\right]|\mathcal{D}_t\right] + \mathrm{E}\left[\widehat{\mathrm{Var}}\left[\boldsymbol{\theta}_t|, \lambda_t, \mathcal{D}_{t-1}\right]|\mathcal{D}_t\right] \\
&= \mathrm{Var}\left[\mathbf{a}_t + \mathsf{R}_t\mathsf{F}\frac{(\lambda_t - f_t)}{q_t}\right] + \mathrm{E}\left[\mathsf{R}_t - \frac{\mathsf{R}_t\mathsf{F}\mathsf{F}^T\mathsf{R}_t}{q_t}|\mathcal{D}_t\right] \\
&= \mathsf{R}_t\mathsf{F}\frac{1}{q_t}\mathrm{Var}\left[\lambda_t|\mathcal{D}_t\right]\frac{1}{q_t}\mathsf{F}^T\mathsf{R}_t + \mathsf{R}_t - \mathsf{R}_t\mathsf{F}\frac{1}{q_t}\mathsf{F}^T\mathsf{R}_t \\
&= \mathsf{R}_t - \mathsf{R}_t\mathsf{F}\left(1 - \frac{\mathrm{Var}\left[\lambda_t|\mathcal{D}_t\right]}{q_t}\right)\frac{1}{q_t}\mathsf{F}^T\mathsf{R}_t
\end{aligned}$$

Following West & Harrison (1997), we use the posterior $\lambda_t|\mathcal{D}_t$ in (5.3) to achieve the result

$$\begin{aligned}
\boldsymbol{\theta}_t|\mathcal{D}_t &\sim [\mathbf{m}_t, \mathsf{C}_t] \\
&= \left[\mathbf{a}_t + \mathsf{R}_t\mathsf{F}\frac{1}{q_t}\left(\underbrace{\mathrm{E}\left[\lambda_t|\mathcal{D}_t\right]}_{f_t^\star} - f_t\right), \mathsf{R}_t - \mathsf{R}_t\mathsf{F}\left(1 - \frac{\overbrace{\mathrm{Var}\left[\lambda_t|\mathcal{D}_t\right]}^{q_t^\star}}{q_t}\right)\frac{1}{q_t}\mathsf{F}^T\mathsf{R}_t\right] \\
&= \left[\mathbf{a}_t + \mathsf{R}_t\mathsf{F}\frac{1}{q_t}\left(f_t^\star - f_t\right), \mathsf{R}_t - \mathsf{R}_t\mathsf{F}\left(1 - \frac{q_t^\star}{q_t}\right)\frac{1}{q_t}\mathsf{F}^T\mathsf{R}_t\right]
\end{aligned} \tag{5.8}$$

Two things to note are that this is an approximation to the moments of the posterior (since $\boldsymbol{\theta}_t|\lambda_t, \mathcal{D}_{t-1}$ is an approximation) and the these calculations simplify when the ca-

[2] For a complete derivation, *cf.* West & Harrison (1997), Section 4.9.2.

nonical link is used, since $\lambda_t = \eta_t$.

A general algorithm for conjugate filtering is presented in Algorithm 5.1.

Algorithm 5.1 Conjugate filtering

---

initialisation

$\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0)$

for $t \leftarrow 1$ to N_{obs}

calculate state prior

$$\begin{aligned} \boldsymbol{\theta}_t|\mathcal{D}_{t-1} &\sim [\mathbf{a}_t, \mathsf{R}_t] \\ \mathbf{a}_t &= \mathsf{G}\mathbf{m}_{t-1} \\ \mathsf{R}_t &= \mathsf{G}\mathsf{C}_{t-1}\mathsf{G}^T + \mathsf{W} \end{aligned}$$

calculate natural parameter prior

$$\begin{aligned} \lambda_t|\mathcal{D}_{t-1} &\sim [f_t, q_t] \\ f_t &= \mathsf{F}^T\mathbf{a}_t \\ q_t &= \mathsf{F}^T\mathsf{R}_t\mathsf{F} \end{aligned}$$

calculate using the conjugate prior $CP(r_t, s_t)$ on η_t (observation distribution dependent), calculate natural parameter posterior

$$\eta_t|\mathcal{D}_t \sim CP(r_t + y_t, s_t + 1)$$

calculate natural parameter posterior moments

$$\lambda_t|\mathcal{D}_t \sim [f_t^\star, q_t^\star]$$

calculate state's partial posterior

$$\begin{aligned} \boldsymbol{\theta}_t|\mathcal{D}_t &\sim [\mathbf{m}_t, \mathsf{C}_t] \\ \mathbf{m}_t &= \mathbf{a}_t + \mathsf{R}_t\mathsf{F}\frac{(f_t^\star - f_t)}{q_t} \\ C_t &= \mathsf{R}_t - \mathsf{R}_t\mathsf{F}\left(1 - \frac{q_t^\star}{q_t}\right)\frac{1}{q_t}\mathsf{F}^T\mathsf{R}_t \end{aligned}$$

5.2 Natural parameter conjugate update

If we consider canonical definition of a DGLM as presented in West & Harrison (1997):

$$p\left(y_t|\eta_t\right) = \exp\left\{\frac{z\left(y_t\right)\eta_t - b\left(\eta_t\right)}{a\left(\phi_t\right)}\right\} c^{\star}\left(y_t\right)$$
$$c^{\star}\left(y_t\right) = \exp c\left(y_t\right),$$

we can see that (in Figure 2.2), in the filtering context, the only unknown quantity is η_t and the prior $p\left(\eta_t|\mathcal{D}_{t-1}\right)$. According to West & Harrison (1997), we can then express the one-step forecast of $y_t|\mathcal{D}_{t-1}$ as

$$p\left(y_t|\mathcal{D}_{t-1}\right) = \int p\left(y_t|\eta_t, \mathcal{D}_{t-1}\right) p\left(\eta_t|\mathcal{D}_{t-1}\right) d\eta_t. \tag{5.9}$$

After we observe y_t we can then use Bayes' Theorem to write the posterior of the natural parameter given the current observation as

$$p\left(\eta_t|\mathcal{D}_t\right) = \frac{p\left(\eta_t|\mathcal{D}_{t-1}\right) p\left(y_t|\eta_t, \mathcal{D}_{t-1}\right)}{\int p\left(\eta_t|\mathcal{D}_{t-1}\right) p\left(y_t|\eta_t, \mathcal{D}_{t-1}\right) d\eta_t}. \tag{5.10}$$

According to West & Harrison (1997), (5.9) and (5.10) are analytically available in the DGLM when the prior belongs to the *conjugate family*. One of the main reasonings of conjugate filtering is that the prior for the state is partially specified by the first (4.5) and second (4.6) moments. However, if r_t and s_t are calculated, the parameters for a conjugate prior (CP) for the natural parameter η_t take the form

$$\begin{aligned} p\left(\eta_t|\mathcal{D}_{t-1}\right) &= c\left(r_t, s_t\right) \exp\left\{r_t\eta_t - s_t b\left(\eta_t\right)\right\} \\ &= \mathrm{CP}\left(r_t, s_t\right) \end{aligned} \tag{5.11}$$

where r_t and $s_t > 0$ are known functions depending on $\mathcal{D}_{t-1}$ and $c\left(\cdot\right)$ works as a normalising function

$$c\left(r_t, s_t\right)^{-1} = \int \exp\left\{r_t\eta_t - s_t a\left(\eta_t\right)\right\} d\eta_t, \tag{5.12}$$

then we could obtain the parameters of the conjugate posterior given the observed data. The next step is to establish the conjugate prior and posterior relations. To do so, we can

start by rewriting the η_t prior (5.11), by assuming[3] $x_t = r_t/s_t$, we have

$$\begin{aligned} p(\eta_t|\mathcal{D}_{t-1}) &= c(r_t, s_t) \exp\{r_t\eta_t - s_t b(\eta_t)\} \\ &= c(r_t, s_t) \exp\{x_t s_t \eta_t - s_t b(\eta_t)\} \\ &= c(r_t, s_t) \exp\{s_t [x_t\eta_t - b(\eta_t)]\}. \end{aligned}$$

If we assume that r_t, s_t and c are known, and using the normalising function in (5.12), we can calculate the one-step ahead forecast in (5.9) as

$$\begin{aligned} p(y_t|\mathcal{D}_{t-1}) &= \int p(y_t|\eta_t, \mathcal{D}_{t-1})\, p(\eta_t|\mathcal{D}_{t-1})\, d\eta_t \\ &= \int \underbrace{\exp\{\phi_t [y_t\eta_t - b(\eta_t)]\}\, c^\star(y_t)}_{p(y_t|\eta_t, \mathcal{D}_{t-1})} \underbrace{c(r_t, s_t) \exp\{r_t\eta_t - s_t b(\eta_t)\}}_{p(\eta_t|\mathcal{D}_{t-1})} d\eta_t \\ &= c^\star(y_t)\, c(r_t, s_t) \int \exp\{\phi_t [y_t\eta_t - b(\eta_t)]\} \exp\{r_t\eta_t - s_t b(\eta_t)\}\, d\eta_t \\ &= c^\star(y_t)\, c(r_t, s_t) \int \exp\{\phi_t y_t\eta_t - \phi_t b(\eta_t) + r_t\eta_t - s_t b(\eta_t)\}\, d\eta_t \\ &= c^\star(y_t)\, c(r_t, s_t) \underbrace{\int \exp\{(r_t + \phi_t y_t)\eta_t - (s_t + \phi_t) b(\eta_t)\}\, d\eta_t}_{c(r_t+\phi_t y_t, s_t+\phi_t)^{-1}} \\ &= \frac{c^\star(y_t)\, c(r_t, s_t)}{c(r_t + \phi_t y_t, s_t + \phi_t)}. \end{aligned} \tag{5.13}$$

Using the expression in (5.14) we can then apply substitution and write the natural parameter posterior as

$$\begin{aligned} p(\eta_t|\mathcal{D}_t) &= \frac{p(\eta_t|\mathcal{D}_{t-1})\, p(y_t|\eta_t, \mathcal{D}_{t-1})}{\int p(\eta_t|\mathcal{D}_{t-1})\, p(y_t|\eta_t, \mathcal{D}_{t-1})\, d\eta_t} \\ &= \frac{\overbrace{c(r_t, s_t) \exp\{r_t\eta_t - s_t b(\eta_t)\}}^{p(\eta_t|\mathcal{D}_{t-1})} \overbrace{\exp\{\phi_t [y_t\eta_t - b(\eta_t)]\}\, c^\star(y_t)}^{p(y_t|\eta_t, \mathcal{D}_{t-1})}}{\underbrace{c^\star(y_t)\, c(r_t, s_t)\, c(r_t + \phi_t y_t, s_t + \phi_t)^{-1}}_{p(y_t|\mathcal{D}_{t-1})}} \\ &= c(r_t + \phi_t y_t, s_t + \phi_t) \exp\{r_t\eta_t - s_t b(\eta_t)\} \exp\{\phi_t [y_t\eta_t - b(\eta_t)]\} \\ &= c(r_t + \phi_t y_t, s_t + \phi_t) \exp\{(r_t + \phi_t y_t)\eta_t - (s_t + \phi_t) b(\eta_t)\}. \end{aligned} \tag{5.14}$$

It is clear that the posterior in (5.14) has the same for of the prior in (5.11). This is one of the conditions to establish a conjugate update, which will happen with a parameter update

[3] As noted in West & Harrison (1997), b being convex means that the prior is unimodal with mode $x_t = b'(\eta_t)$.

of

$$\begin{aligned} r_{t+1} &= r_t + \phi_t y_t \\ s_{t+1} &= s_t + \phi_t. \end{aligned}$$

In the following sections we will describe the standard conjugate scheme as derived in West & Harrison (1997) for each of the DGLMs considered with the conjugate relations summarised in Table 5.1 and prior and posterior moments of the natural parameter summarised in Table 5.2.

5.2.1 Binomial

For the BDLM, as described in Section 2.3.3, West & Harrison (1997) defines the $p_t = {}^{e^{\eta_t}}/_{1+e^{\eta_t}}$ prior to be beta, that is

$$\begin{aligned} p\left(p_t|\mathcal{D}_{t-1}\right) &\sim \mathrm{B}\left(r_t, s_t\right) \\ &= c\left(r_t, s_t\right) p_t^{r_t-1}\left(1-p_t\right)^{s_t-1}, \qquad 0 \leq p_t \leq 1 \end{aligned}$$

with the normalising constant $c\left(\cdot,\cdot\right)$, as defined in (5.12), as

$$c\left(r_t, s_t\right) = \frac{\Gamma\left(r_t + s_t\right)}{\Gamma\left(r_t\right)\Gamma\left(s_t\right)}.$$

By considering the *digamma* function

$$\gamma\left(x\right) = \frac{\Gamma'\left(x\right)}{\Gamma\left(x\right)},$$

using the natural parameter $\eta_t = \log\left\{{}^{p_t}/_{(1-p_t)}\right\}$ allows to write (5.11) as

$$p\left(\eta_t|\mathcal{D}_{t-1}\right) = c\left(r_t, s_t\right)\exp\left\{r_t\eta_t - s_t b\left(\eta_t\right)\right\}$$

West & Harrison (1997) states that we can write

$$\begin{aligned} \mathrm{E}\left[\eta_t|\mathcal{D}_{t-1}\right] &= f_t = \gamma\left(r_t\right) - \gamma\left(s_t\right) \\ \mathrm{Var}\left[\eta_t|\mathcal{D}_{t-1}\right] &= q_t = \gamma'\left(r_t\right) - \gamma'\left(s_t\right) \end{aligned}$$

Applying an approximation to the digamma function[4] such that

$$\begin{aligned}\gamma(x) &\approx \log(x)\\ \gamma'(x) &\approx \frac{1}{x},\end{aligned}$$

we can approximate moments of natural parameter prior by

$$\begin{aligned}\mathrm{E}[\eta_t|\mathcal{D}_{t-1}] &= f_t\\ &= \gamma(r_t) - \gamma(s_t)\\ &\approx \log(r_t) - \log(s_t) = \log\left(\frac{r_t}{s_t}\right)\\ \mathrm{Var}[\eta_t|\mathcal{D}_{t-1}] &= q_t\\ &= \gamma'(r_t) - \gamma'(s_t)\\ &\approx \frac{1}{r_t} + \frac{1}{s_t},\end{aligned}$$

which we can rearranged resulting in

$$\begin{aligned}r_t &= \frac{1+\exp(f_t)}{q_t}\\ s_t &= \frac{1+\exp(-f_t)}{q_t}.\end{aligned}$$

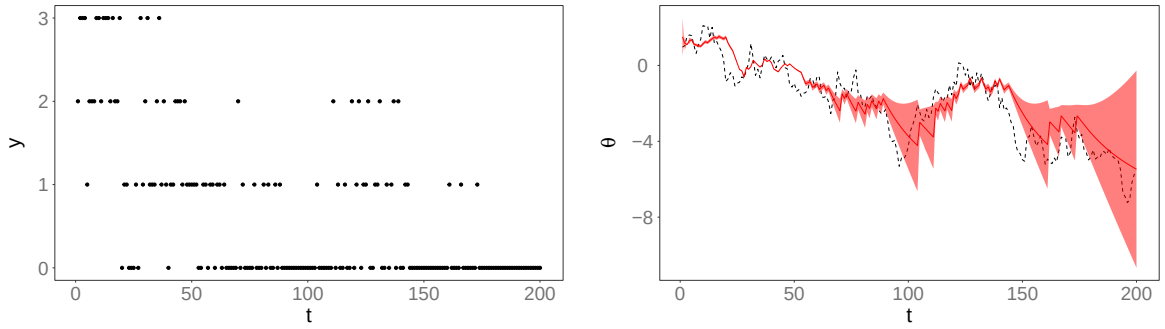
Having the values of r_t and s_t the p_t posterior can be updated as

$$p(p_t|\mathcal{D}_t) \sim \mathrm{B}(r_t + y_t, s_t + n_t - y_t).$$

Using the same reasoning as for the prior, we get the following moments

$$\begin{aligned}\mathrm{E}[\eta_t|\mathcal{D}_t] &= f_t^\star\\ &= \gamma(r_t + y_t) - \gamma(s_t + n_t - y_t)\\ &\approx \log(r_t + y_t) - \log(s_t + n_t - y_t)\\ &= \log\left\{\frac{r_t + y_t}{s_t + n_t - y_t}\right\}\\ \mathrm{Var}[\eta_t|\mathcal{D}_t] &= q_t^\star\\ &= \gamma'(r_t + y_t) - \gamma'(s_t + n_t - y_t)\\ &\approx \frac{1}{r_t + y_t} - \frac{1}{s_t + n_t - y_t}\end{aligned}$$

[4] *cf.* Appendix F.5



(a) $y_{1:t}$ (b) $\theta_{0:t}$

Figure 5.1: Observations (*left*) and state estimation (*right*) for a $\mathcal{M} = \{\mathcal{P}(1)\}$ Binomial DLM with $\boldsymbol{\Phi} = \{\tau^2, k\} = \{0.25, 3\}$ using CF. Dashed line represents the model's realisation true state, solid red line the filtering density mean and shaded area the 90% CI.

Example. Conjugate filtering for a $\mathcal{M} = \{\mathcal{P}(1)\}$ BDLM.

Here we consider a $\mathcal{M} = \{\mathcal{P}(1)\}$ BDLM with parameters $\boldsymbol{\Phi} = \{\tau^2, k\} = \{0.25, 3\}$, where k is the number of trials/categories, corresponding to the formulation

$$\begin{aligned} y_t|\theta_t, \boldsymbol{\Phi} &\sim \text{Binom}(\eta_t, 3) \\ \eta_t|\theta_t &= \text{logit}^{-1}(\theta_t) \\ \theta_t|\boldsymbol{\Phi} &\sim \mathcal{N}(\theta_{t-1}, 0.25). \end{aligned}$$

The state prior used was $\theta_0 \sim \mathcal{N}(1.5, 10)$ and state estimation was performed using the CF technique. The data for a realisation of this model and respective true and estimated state can be viewed in Figure 5.1.

5.2.2 Poisson

For the PoDLM, as defined in Section 2.3.2, according to West *et al.* (1985), the mean prior for μ_t will be

$$p(\mu_t|\mathcal{D}_{t-1}) \sim \mathcal{G}(r_t, s_t)$$

In this case the parameters of the conjugate prior, according to West *et al.* (1985), can be written (and using the digamma approximation) as

$$\begin{aligned} \mathrm{E}\left[\eta_t|\mathcal{D}_{t-1}\right] &= f_t = \gamma\left(r_t\right) - \log\left(s_t\right) \\ &\approx \log\left(r_t\right) - \log\left(s_t\right) \\ \mathrm{Var}\left[\eta_t|\mathcal{D}_{t-1}\right] &= q_t = \gamma'\left(r_t\right) \\ &\approx \frac{1}{r_t}. \end{aligned}$$

Rearranging we get in terms of f_t and q_t, we have

$$r_t = \frac{1}{q_t} \tag{5.15}$$

$$s_t = \frac{\exp\left(-f_t\right)}{q_t}. \tag{5.16}$$

The posterior update will be in the form

$$p\left(\mu_t|\mathcal{D}_t\right) \sim \mathcal{G}\left(r_t + y_t, s_t + 1\right),$$

and posterior can the be calculated from

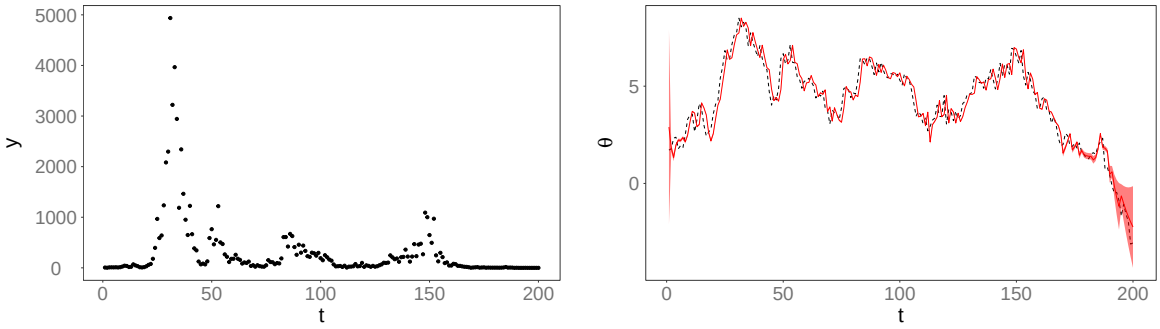
$$\begin{aligned} \mathrm{E}\left[\eta_t|\mathcal{D}_t\right] &= f_t^\star \\ &= \gamma\left(r_t + y_t\right) - \log\left(s_t + 1\right) \\ &\approx \log\left(r_t + y_t\right) - \log\left(s_t + 1\right) \\ \mathrm{Var}\left[\eta_t|\mathcal{D}_t\right] &= q_t^\star \\ &= \gamma'\left(r_t + y_t\right) \\ &\approx \frac{1}{r_t + y_t}. \end{aligned}$$

Example. Conjugate filtering for a $\mathcal{M} = \{\mathcal{P}\left(1\right)\}$ PoDLM.

Here we consider a $\mathcal{M} = \{\mathcal{P}\left(1\right)\}$ PoDLM with parameters $\boldsymbol{\Phi} = \left\{\tau^2\right\} = \{0.3\}$, corresponding to the formulation

$$\begin{aligned} y_t|\theta_t, \boldsymbol{\Phi} &\sim \mathrm{Po}\left(e^{\eta_t}\right) \\ \eta_t &= \theta_t \\ \theta_t|\boldsymbol{\Phi} &\sim \mathcal{N}\left(\theta_t, 0.3\right). \end{aligned}$$

The state prior used was $\theta_0 \sim \mathcal{N}\left(2.9, 10\right)$ and state estimation was performed using the CF technique. The data for a realisation of this model and respective true and estimated



(a) $y_{1:t}$ (b) $\theta_{0:t}$

Figure 5.2: Observations (*left*) and filtering density estimation mean (*right*) for a $\mathcal{M} = \{\mathcal{P}(1)\}$ PoDLM with $\boldsymbol{\Phi} = \{\tau^2\} = \{0.3\}$ using CF. Dashed line represents the model's realisation true state, colour line the filtering density mean and shaded area 90% CI.

state can be viewed in Figure 5.2.

5.2.3 Normal

Although a moment approximation is not needed for the NDLM, since we have the exact solution, for completeness the conjugate filtering derivation is also presented. The conjugate prior of the natural parameter will be

$$\begin{aligned} \mathrm{E}\left[\eta_t|\mathcal{D}_{t-1}\right] &= f_t \\ &= \frac{r_t}{s_t} \\ \mathrm{Var}\left[\eta_t|\mathcal{D}_{t-1}\right] &= q_t \\ &= s_t^{-1} \end{aligned}$$

Rearranging in terms of r_t and s_t we have

$$\begin{aligned} r_t &= \frac{f_t}{q_t} \\ s_t &= \frac{1}{q_t}. \end{aligned}$$

As the the forecast density will also be normal as we've seen in (4.9) we will have

$$\begin{aligned} \mathrm{E}\left[y_t|\mathcal{D}_{t-1}\right] &= \frac{r_t}{s_t} \\ \mathrm{Var}\left[y_t|\mathcal{D}_{t-1}\right] &= V + s_t^{-1}. \end{aligned}$$

The natural parameter posterior, since $\eta_t = \mu_t$, will have (and considering that $\phi_t = V^{-1}$ in the NDLM)

$$\begin{aligned}
\mathrm{E}\left[\eta_t|\mathcal{D}_t\right] &= f_t^\star \\
&= \frac{r_t}{s_t} + \frac{\phi_t}{s_t+\phi_t}\left(y_t - \frac{r_t}{s_t}\right) \\
&= \frac{r_t}{s_t} + \frac{V^{-1}}{s_t+V^{-1}}\left(y_t - \frac{r_t}{s_t}\right) \\
&= \frac{r_t V + y_t}{s_t V + 1} \\
\mathrm{Var}\left[\eta_t|\mathcal{D}_{t-1}\right] &= q_t^\star \\
&= s_t^{-1} - \left(\frac{\phi_t}{s_t+\phi_t}\right)^2\left(V + s_t^{-1}\right) \\
&= s_t^{-1} - \left(\frac{V^{-1}}{s_t+V^{-1}}\right)^2\left(V + s_t^{-1}\right) \\
&= \frac{V}{s_t V + 1}.
\end{aligned}$$

5.3 Forecasting

Considering again the DGLM using the canonical formulation of (2.5), we assume that at time t, the starting point of the forecast, we have the posterior moments of the state vector

$$\boldsymbol{\theta}_t|\mathcal{D}_t \sim \left[\mathbf{m}_t, \mathsf{C}_t\right].$$

Using the state forecast KF recursions as described in (4.14) and (4.15) we have write the partially specified propagation density

$$\boldsymbol{\theta}_{t+k}|\mathcal{D}_t \sim \left[\mathbf{a}_t\left(k\right), \mathsf{R}_t\left(k\right)\right].$$

The predictive forecast moments for the signal $\lambda_{t+k} = \mathsf{F}^T\boldsymbol{\theta}_{t+k}$ can be specified based on the KF recursions in (4.16) and (4.17) such that

$$\lambda_{t+k}|\mathcal{D}_t \sim \left[f_t\left(k\right), q_t\left(k\right)\right],$$

with moments updated in the non-linear DGLM case as

$$\begin{aligned}
f_t\left(k\right) &= \mathsf{F}^T\mathbf{a}_t\left(k\right) \\
q_t\left(k\right) &= \mathsf{F}^T\mathsf{R}_t\left(k\right)\mathsf{F}.
\end{aligned}$$

As stated by West & Harrison (1997), assuming we have the conjugate prior for $\eta_{t+k} = g^{-1}(\lambda_{t+k})$, we can define the k-step ahead observation forecast base on the formulation already provided in (5.13) for the one-step ahead forecast, that is

$$\begin{aligned} p(y_{t+k}|\mathcal{D}_t) &= \int p(y_{t+k}|\eta_{t+k}, \mathcal{D}_t)\, p(\eta_{t+k}|\mathcal{D}_t)\, d\eta_{t+k} \\ &= \frac{c^\star(y_{t+k}, \phi_{t+k})\, c(r_t(k), s_t(k))}{c(r_t(k) + \phi_{t+k} y_{t+k}, s_t(k) + \phi_{t+k})}, \end{aligned}$$

where $r_t(k)$ and $s_t(k)$ are evaluated from $f_t(k)$ and $q_t(k)$, the mean and variance of $g(\eta_{t+k})|\mathcal{D}_t$, that is

$$\begin{aligned} f_t(k) &= \mathrm{E}[g(\eta_{t+k})|\mathcal{D}_t] \\ q_t(k) &= \mathrm{E}[g(\eta_{t+k})|\mathcal{D}_t], \end{aligned}$$

and the distribution of $\eta_{t+k}|\mathcal{D}_t$ can now be calculated from the approximations in Section 5.2 as we will see in the following section, along with some forecasting examples.

5.3.1 Poisson DLM

Using the previously calculated values r_t and s_t in for a PoDLM, respectively in (5.15) and (5.16), that is

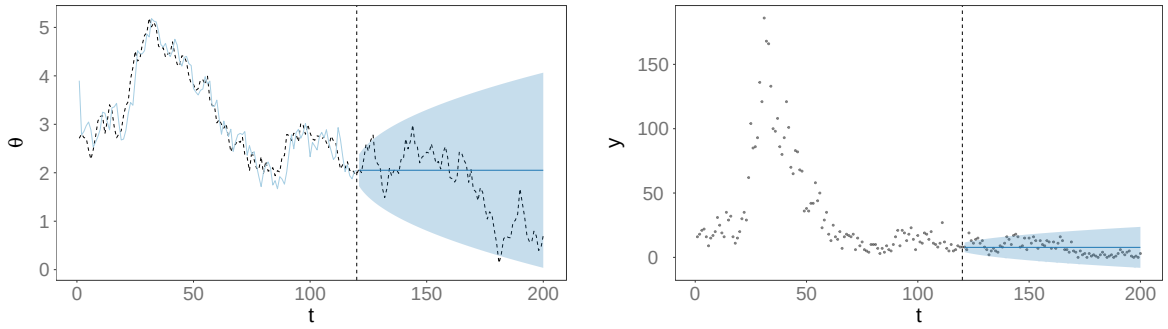
$$\begin{aligned} r_t &= \frac{1}{q_t} \\ s_t &= \frac{\exp(-f_t)}{q_t}, \end{aligned}$$

and defining $r_t(k)$ and $s_t(k)$ according to $f_t(k)$ and $q_t(k)$ the k-step ahead forecast distribution of $y_{t+k}|\mathcal{D}_t$ is

$$p(y_{t+k}|\mathcal{D}_t) = \binom{r_t(k) + y_{t+k} - 1}{y_{t+k}} \left(\frac{s_t(k)}{1 + s_t(k)}\right)^{r_t(k)} \left(\frac{1}{1 + s_t(k)}\right)^{y_{t+k}},$$

which is a negative binomial distribution. The forecast mean and variance can be calculated using conditional expectations, *i.e.*

$$\begin{aligned} y_t(k) &= \mathrm{E}[y_{t+k}|\mathcal{D}_t] \\ &= \mathrm{E}[\mathrm{E}[y_{t+k}|\lambda_{t+k}]\,|\mathcal{D}_t] \\ &= \frac{r_t(k)}{s_t(k)}, \end{aligned}$$



(a) $\hat{\theta}_{t:t+k}$ for $t = 120, k = 80$

(b) $\hat{y}_{t:t+k}$ for $t = 120, k = 80$

Figure 5.3: State filtering density and k-step ahead forecast (*left*, dashed line represents "true" value from the realisation, colour line the filtering density mean and shaded area the 90% CI) and observation (*right*, colour line represents forecast mean and shaded area the 90% CI) k-step ahead forecasts for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM using the KF.

and

$$\begin{aligned} \text{Var}\left[y_{t+k}|\mathcal{D}_t\right] &= \text{E}\left[\text{Var}\left[y_{t+k}|\lambda_{t+k}\right]|\mathcal{D}_t\right] + \text{Var}\left[\text{E}\left[y_{t+k}|\lambda_{t+k}\right]|\mathcal{D}_t\right] \\ &= \frac{r_t(k)\left[s_t(k)+1\right]}{\left(s_t(k)\right)^2} \end{aligned}$$

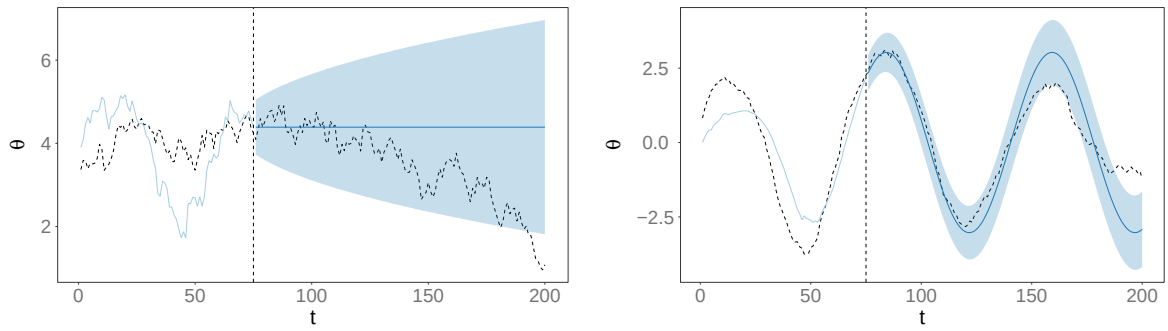
Example. State and observation forecast in a $\mathcal{M} = \{\mathcal{P}(1)\}$ PoDLM.

Here we present state and observation k-step ahead forecasts for a $N_{obs} = 200$ realisation of a $\mathcal{M} = \{\mathcal{P}(1)\}$ PoDLM using the CF method. The parameter set used was $\Phi = \{\tau^2\} = \{0.05\}$. The state prior used was $\theta_0 \sim \mathcal{N}(3.9, 5)$. In Figure 5.3a we can see the k-step ahead state forecast and in Figure 5.3b the k-step observation forecast, both for $k = 80$ and starting from $t = 120$.

Example. State and observation forecast in $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(75, 1)\}$ PoDLM.

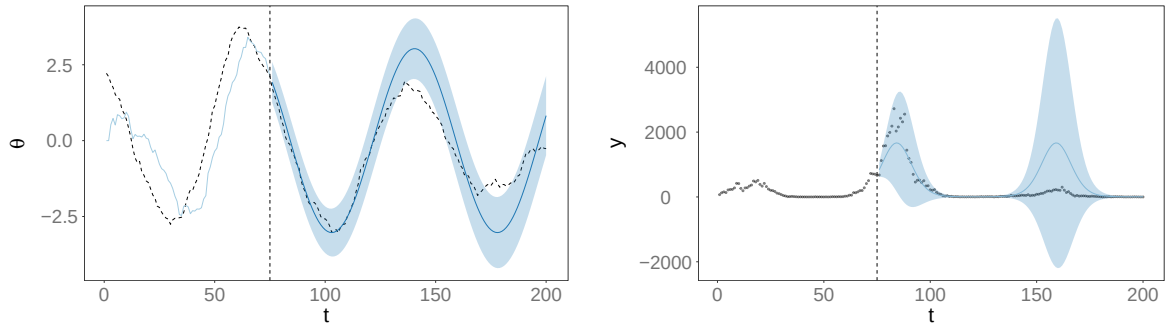
Here we consider state and observation k-step ahead forecasts for a $N_{obs} = 200$ realisation of a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(75, 1)\}$ PoDLM using the CF method. The parameter set used was

$$\Phi = \{\mathsf{W}\} = \left\{ \begin{bmatrix} 0.05 & 0 & 0 \\ 0 & 0.01 & 0 \\ 0 & 0 & 0.01 \end{bmatrix} \right\},$$



(a) $\hat{\boldsymbol{\theta}}^1_{t:t+k}$ for $t = 75, k = 125$

(b) $\hat{\boldsymbol{\theta}}^2_{t:t+k}$ for $t = 75, k = 125$



(c) $\hat{\boldsymbol{\theta}}^3_{t:t+k}$ for $t = 75, k = 125$

(d) $y_{t:t+k}$ for $t = 75, k = 125$

Figure 5.4: Observation (*bottom right*, colour line represents the forecast mean and shaded area the 90% CI) and state (dashed line represents the "true" values from the realisation, colour line the forecast mean and shaded area the 90% CI) k-step ahead forecasts for a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(75, 1)\}$ PoDLM using CF.

with a state prior

$$\theta_0 \sim \mathcal{N}\left(\begin{bmatrix} 3.9 & 0 & 0 \end{bmatrix}^T, \begin{bmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix}\right).$$

In Figures 5.4a, 5.4b and 5.4c we can see the k-step ahead state forecast for each state vector components and in Figure 5.4d the k-step ahead observation forecast, all of them for $k = 125$ and starting from $t = 75$.

5.3.1.1 Binomial DLM

For the BDLM, given y_t, to calculate the k-step ahead forecast we start by assuming that

$$p_{t+k}|\mathcal{D}_t \sim \mathrm{B}\left(r_t(k), s_t(k) - r_t(k)\right).$$

The k-step ahead forecast distribution is then given by

$$\begin{aligned} p\left(y_{t+k}|\mathcal{D}_t\right) &= \frac{\Gamma\left(s_t\left(k\right)\right)}{\Gamma\left(r_t\left(k\right)\right)\Gamma\left(s_t\left(k\right)-r_t\left(k\right)\right)\Gamma\left(s_t\left(k\right)+n_{t+k}\right)} \\ &\quad \times\frac{1}{n_{t+k}}\binom{n_{t+k}}{y_{t+k}}\Gamma\left(r_t\left(k\right)+y_{t+k}\right)\Gamma\left(s_t\left(k\right)-r_t\left(k\right)+n_{t+k}-y_{t+k}\right). \end{aligned}$$

Using conditional expectation we have the forecast mean and variance as

$$\begin{aligned} y_t\left(k\right) &= \mathrm{E}\left[y_{t+k}|\mathcal{D}_t\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[y_{t+k}|p_{t+k}\right]|\mathcal{D}_t\right] \\ &= \frac{n_{t+k}\left[r_t\left(k\right)+1\right]}{r_t\left(k\right)+s_t\left(k\right)+1} \\ \mathrm{Var}\left[y_{t+k}|\mathcal{D}_t\right] &= \mathrm{E}\left[\mathrm{Var}\left[y_{t+k}|p_{t+k}\right]|\mathcal{D}_t\right]+\mathrm{Var}\left[\mathrm{E}\left[y_{t+k}|p_{t+k}\right]\right] \\ &= \frac{n_{t+k}\left[r_t\left(k\right)+1\right]}{r_t\left(k\right)+s_t\left(k\right)+1}-\frac{n_{t+k}\left[r_t\left(k\right)+1\right]\left[r_t\left(k\right)+2\right]}{\left(r_t\left(k\right)+s_t\left(k\right)+1\right)\left(r_t\left(k\right)+s_t\left(k\right)+2\right)} \\ &\quad +\frac{n_{t+k}^2\left[r_t\left(k\right)+1\right]s_t\left(k\right)}{\left(r_t\left(k\right)+s_t\left(k\right)+1\right)^2\left(r_t\left(k\right)+s_t\left(k\right)+2\right)}. \end{aligned}$$

Example. State and observation forecast in $\mathcal{M}=\{\mathcal{P}\left(1\right)\}$ BDLM.

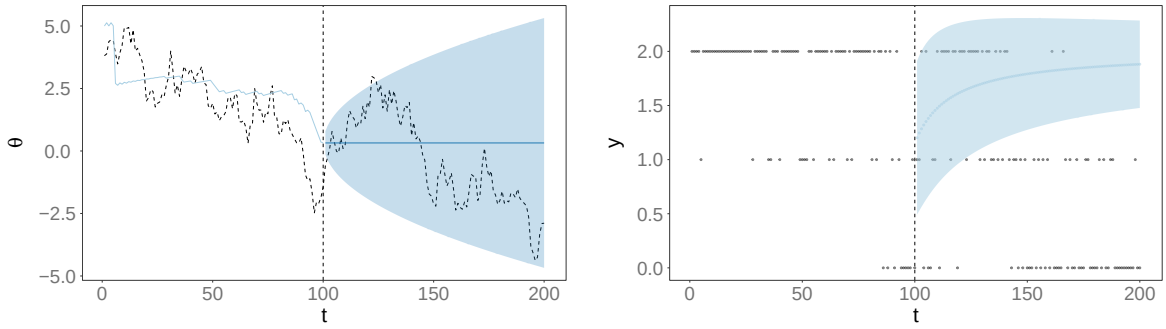
Here we present state and observations k-step ahead forecasts for a $N_{obs}=200$ realisation of a $\mathcal{M}=\{\mathcal{P}\left(1\right)\}$ BDLM in the form

$$\begin{aligned} y_t|\eta_t,\Phi &\sim \mathrm{Binom}\left(\eta_t,2\right). \\ \eta_t|\theta_t &= \mathrm{logit}^{-1}\left(\theta_t\right) \\ \theta_t|\theta_{t-1},\Phi &\sim \mathcal{N}\left(\theta_{t-1},0.25\right), \end{aligned}$$

using the CF method. The parameter set used was $\Phi=\left\{\tau^2\right\}=\{0.25\}$. The state prior used was $\theta_0\sim\mathcal{N}\left(5,5\right)$. In Figure 5.5a we can see the k-step ahead state forecast and in Figure 5.5b the k-step observation forecast, both for for $k=100$ and starting from $t=100$.

5.4 Summary

A summary for the natural parameter's prior and posterior forms for the considered DGLMs is presented in Table 5.1 along with prior and posterior moments approximation in Table 5.2.



(a) $\hat{\theta}_{t:t+k}$ for $t = 100, k = 100$

(b) $\hat{\theta}_{t:t+k}$ for $t = 100, k = 100$

Figure 5.5: State (*left*, dashed line represents the "true" values from the realisation, colour line the forecast mean and shaded area the 90%CI) and observation (*right*, colour line represents the forecast mean and shaded area the 90% CI) k-step ahead forecasts for a $\mathcal{M} = \{\mathcal{P}(1)\}$ BDLM using the CF

DGLM	**Link**	**Prior**	**Posterior**
Binomial	$\log\left(\frac{p_t}{1-p_t}\right)$	$\mathrm{B}(r_t, s_t)$	$\mathrm{B}(r_t + y_t s_t + y_t - n_t)$
Poisson	$\log \lambda_t$	$\mathcal{G}(r_t, s_t)$	$\mathcal{G}(r_t + y_t, s_t + 1)$
Normal	μ_t	$\mathcal{N}\left(\mu_t s_t^{-1}, s_t^{-1}\right)$	$\mathcal{N}\left(\frac{r_t + V^{-1} y_t}{s_t + V^{-1}}, \frac{1}{s_t + V^{-1}}\right)$

Table 5.1: Conjugate update for η_t

DGLM	r_t	s_t	$f_t^\star$	$q_t^\star$
Binomial	$\frac{1+\exp(f_t)}{q_t}$	$\frac{1+\exp(-f_t)}{q_t}$	$\log\left\{\frac{r_t+y_t}{s_t+n_t-y_t}\right\}$	$\frac{1}{r_t+y_t} - \frac{1}{s_t+n_t-y_t}$
Poisson	$\frac{1}{q_t}$	$\frac{\exp(-f_t)}{q_t}$	$\log(r_t + y_t) - \log(s_t + 1)$	$\frac{1}{r_t+y_t}$
Normal	$\frac{f_t}{q_t}$	$\frac{1}{q_t}$	$\frac{r_t V + y_t}{s_t V + 1}$	$\frac{V}{s_t V + 1}$

Table 5.2: Prior and posterior moment approximations for η_t

Chapter 6

Importance Sampling

An additional method to perform online state estimation is Sequential Monte Carlo (SMC). SMC is built on the principle of Importance Sampling (IS) which we will also detail in this chapter.

In this context, we are interested in the estimation of $p(\boldsymbol{\theta}|\mathcal{D}_t)$, the state's posterior probability density[1] given the entirety of the data $\mathcal{D}_t = y_{1:t} = \{y_1, \ldots, y_t\}$. This can be viewed as the calculation of the expectation

$$\bar{g} = \mathrm{E}\left[g(\boldsymbol{\theta})|\mathcal{D}_t\right] = \int g(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D}_t)\, d\boldsymbol{\theta}, \tag{6.1}$$

where $g(\cdot)$ is an arbitrary function. Assuming we could produce samples $\left\{\boldsymbol{\theta}^{(i)}\right\}_{i=1}^{N_p}$ directly from the true posterior $p(\boldsymbol{\theta}|\mathcal{D}_t)$, an empirical (Monte Carlo) approximation of this distribution would be

$$p(\boldsymbol{\theta}|\mathcal{D}_t) \approx \frac{1}{N}\sum_{i=1}^{N_p} \delta\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}\right),$$

where δ is the Dirac δ function. In this case the expectation in (6.1) is

$$\begin{aligned}
\bar{g} &= \int g(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D}_t)\, d\boldsymbol{\theta} \\
&= \int g(\boldsymbol{\theta})\, \frac{1}{N}\sum_{i=1}^{N_p} \delta\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}\right) d\boldsymbol{\theta} \\
&= \frac{1}{N}\sum_{i=1}^{N} g\left(\boldsymbol{\theta}^{(i)}\right).
\end{aligned} \tag{6.2}$$

However, according to Carpenter *et al.* (1999), one of the problems with the calculation

[1] The dependence of $\boldsymbol{\theta}$ on t will be dropped temporarily for simplicity.

in (6.1) is that this distribution may be highly complex and with high dimensionality and typically we cannot sample from it directly. In such cases we can employ a Monte Carlo approximation by producing random samples $\left\{\boldsymbol{\theta}^{(i)}\right\}_{i=1}^{N_p}$ from a *importance density* $\pi(\cdot)$, having a larger support than $p(\cdot)$ (as illustrated in Figure 6.1), from which we can easily sample, such that

$$\boldsymbol{\theta}^{(i)} \sim \pi\left(\boldsymbol{\theta}|\mathcal{D}_t\right), \tag{6.3}$$

and hence

$$\pi\left(\boldsymbol{\theta}|\mathcal{D}_t\right) \approx \frac{1}{N_p}\sum_{i=1}^{N_p}\delta\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(i)}\right).$$

In this case the approximation to (6.1) would then be

$$\begin{aligned}
\bar{g} &= \int g\left(\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}|\mathcal{D}_t\right) d\boldsymbol{\theta} \\
&= \int g\left(\boldsymbol{\theta}\right) \frac{p\left(\boldsymbol{\theta}^{(i)}|\mathcal{D}_t\right)}{\pi\left(\boldsymbol{\theta}^{(i)}|\mathcal{D}_t\right)} \pi\left(\boldsymbol{\theta}^{(i)}|\mathcal{D}_t\right) d\boldsymbol{\theta} \\
&\approx \int g\left(\boldsymbol{\theta}\right) \frac{p\left(\boldsymbol{\theta}^{(i)}|\mathcal{D}_t\right)}{\pi\left(\boldsymbol{\theta}^{(i)}|\mathcal{D}_t\right)} \frac{1}{N}\sum_{i=1}^{N_p}\delta\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(i)}\right) d\boldsymbol{\theta} \\
&= \frac{1}{N_p}\sum_{i=1}^{N_p} \underbrace{\frac{p\left(\boldsymbol{\theta}^{(i)}|\mathcal{D}_t\right)}{\pi\left(\boldsymbol{\theta}^{(i)}|\mathcal{D}_t\right)}}_{w^{(i)}} g\left(\boldsymbol{\theta}^{(i)}\right) \\
&= \frac{1}{N_p}\sum_{i=1}^{N_p} w^{(i)} g\left(\boldsymbol{\theta}^{(i)}\right).
\end{aligned} \tag{6.4}$$

In (6.4), the quantities

$$w^{(i)} = \frac{p\left(\boldsymbol{\theta}^{(i)}|\mathcal{D}_t\right)}{\pi\left(\boldsymbol{\theta}^{(i)}|\mathcal{D}_t\right)}$$

for $i = 1, \ldots N_p$ are usually called the *unnormalised importance weights*, which account for the difference between the target distribution $p(\cdot)$ and the importance density $\pi(\cdot)$. We can further define the *normalised importance weights* according to the condition

$$\tilde{w}^{(i)} = \frac{w^{(i)}}{\sum_{i=1}^{N_p} w^{(i)}}, \qquad \sum \tilde{w}^{(i)} = 1.$$

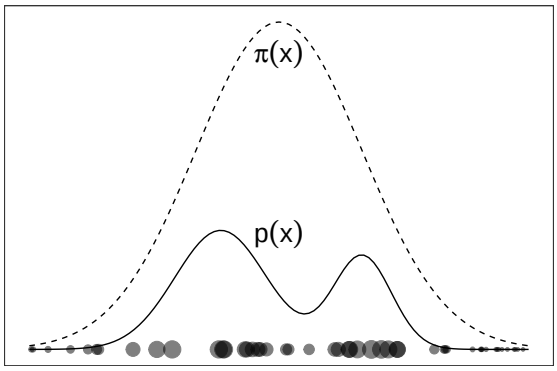


Figure 6.1: Illustration of Importance Sampling with $p(x)$ as the target and $\pi(x)$ as the scaled importance distribution. Circles indicate samples from $\pi(\cdot)$, with size proportional to the weight.

In this case, the approximation in (6.4) could be written as

$$\bar{g} = \sum_{i=1}^{N_p} \tilde{w}^{(i)} g\left(\boldsymbol{\theta}^{(i)}\right).$$

It is important to note that $\sum_{i=1}^{N_p} \tilde{w}^{(i)} g\left(\boldsymbol{\theta}^{(i)}\right) \xrightarrow{p} \int g(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D}_t)\, d\boldsymbol{\theta}$ when $N_p \to \infty$.

Using IS methods we can then specify the approximation of (6.1) by a set of N_p samples $\boldsymbol{\theta}^{(i)}$ along with their corresponding normalised importance weights $\tilde{w}^{(i)}$, that is, by

$$\left\{\boldsymbol{\theta}^{(i)}, \tilde{w}^{(i)}\right\}_{i=1}^{N_p}. \tag{6.5}$$

As noted in Tokdar & Kass (2010), if the variance of (6.1) exists, it can be calculated by (using the result in (6.2))

$$\begin{aligned}
\mathrm{Var}\left[\bar{g}\right] &= \mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N} g\left(\boldsymbol{\theta}^{(i)}\right)\right] \\
&= \frac{1}{N}\mathrm{Var}\left[g(\boldsymbol{\theta})\right] \\
&= \frac{1}{N}\int \left[g(\boldsymbol{\theta}) - \mathrm{E}\left[g(\boldsymbol{\theta})\right]\right]^2 p(\boldsymbol{\theta}|\mathcal{D}_t)\, d\boldsymbol{\theta}.
\end{aligned} \tag{6.6}$$

The choice of importance proposal is therefore crucial to minimise the variance of the Monte Carlo estimation. In Section 6.2 we will look at different importance proposals and specifically the *optimal importance density*, which minimises the variance in (6.6).

6.1 IS for filtering problems

In the previous section we have looked at how to estimate $p(\boldsymbol{\theta}|\mathcal{D}_t)$ using IS methods, however, in the context of this thesis, we are interested in applying the IS method to the sequential estimation of the state, that is $p(\boldsymbol{\theta}_{0:t}|\mathcal{D}_t)$ as the observations $\mathcal{D}_t = \{y_1, y_2, \ldots, y_t\}$ arrive. At this point we introduce the notation Θ_t to denote the sequence of state vectors up until time t, that is

$$\Theta_t = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t\}.$$

Basically we are now interested in the calculation of an expectation of the form of (6.1) for

$$\bar{g} = \mathrm{E}[\Theta_t|\mathcal{D}_t] = \int g(\Theta_t)\, p(\Theta_t|\mathcal{D}_t)\, d\Theta_t. \tag{6.7}$$

Considering that the full posterior for $p(\Theta_t|\mathcal{D}_t)$ can be written, according to (3.2), as

$$p(\Theta_t|\mathcal{D}_t) = \frac{p(\Theta_t, \mathcal{D}_t)}{p(\mathcal{D}_t)},$$

taking $p(\mathcal{D}_t)$ as a normalisation constant, allows us to write

$$p(\Theta_t|\mathcal{D}_t) \propto p(\Theta_t, \mathcal{D}_t).$$

Taking into account the Markovian properties of the DGLM as detailed in Section 2.1, we can decompose the above posterior as in (3.8), that is

$$p(\Theta_t, \mathcal{D}_t) = p(\Theta_{t-1}, \mathcal{D}_{t-1})\ \underbrace{p(y_t|\boldsymbol{\theta}_t)}_{\text{measurement}}\ \underbrace{p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})}_{\text{system}}.$$

With this factorisation, and taking into account

$$p(\mathcal{D}_t) = \int p(\Theta_t, \mathcal{D}_t)\, d\Theta_t,$$

we can then rewrite the IS sampling expression in (6.7) as

$$\begin{aligned}
\bar{g} &= \int g(\Theta_t)\, p(\Theta_t|\mathcal{D}_t)\, d\Theta_t \\
&= \int g(\Theta_t)\, \frac{p(\Theta_t, \mathcal{D}_t)}{p(\mathcal{D}_t)} d\Theta_t \\
&= \frac{\int g(\Theta_t)\, p(\Theta_t, \mathcal{D}_t)\, d\Theta_t}{\int p(\Theta_t, \mathcal{D}_t)\, d\Theta_t}.
\end{aligned}$$

If we now assume an *importance density*, $\pi(\cdot)$, such that we can sample the state vector *sequence*, that is $\Theta_t^{(i)} = \left\{\boldsymbol{\theta}_0^{(i)}, \boldsymbol{\theta}_1^{(i)}, \ldots, \boldsymbol{\theta}_t^{(i)}\right\}$, given the data such as

$$\Theta_t^{(i)} \sim \pi\left(\Theta_t|\mathcal{D}_t\right), \qquad i = 1, \ldots, N_p$$

we could then can apply the IS method of the previous section to the estimation of (6.7):

$$\begin{aligned}
\bar{g} &= \frac{\int g\left(\Theta_t\right) p\left(\Theta_t, \mathcal{D}_t\right) d\Theta_t}{\int p\left(\Theta_t, \mathcal{D}_t\right) d\Theta_t} \\
&= \int g\left(\Theta_t\right) \frac{\pi\left(\Theta_t|\mathcal{D}_t\right)}{\pi\left(\Theta_t|\mathcal{D}_t\right)} p\left(\Theta_t, \mathcal{D}_t\right) d\Theta_t \left[\int \frac{\pi\left(\Theta_t|\mathcal{D}_t\right)}{\pi\left(\Theta_t|\mathcal{D}_t\right)} p\left(\Theta_t, \mathcal{D}_t\right) d\Theta_t\right]^{-1} \\
&= \int g\left(\Theta_t\right) \frac{p\left(\Theta_t, \mathcal{D}_t\right)}{\pi\left(\Theta_t|\mathcal{D}_t\right)} \pi\left(\Theta_t|\mathcal{D}_t\right) d\Theta_t \left[\int \frac{p\left(\Theta_t, \mathcal{D}_t\right)}{\pi\left(\Theta_t|\mathcal{D}_t\right)} \pi\left(\Theta_t|\mathcal{D}_t\right) d\Theta_t\right]^{-1} \\
&= \mathrm{E}\left[g\left(\Theta_t\right) \frac{p\left(\Theta_t, \mathcal{D}_t\right)}{\pi\left(\Theta_t|\mathcal{D}_t\right)}\right] \left\{\mathrm{E}\left[\frac{p\left(\Theta_t, \mathcal{D}_t\right)}{\pi\left(\Theta_t|\mathcal{D}_t\right)}\right]\right\}^{-1} \\
&= \frac{1}{N_p} \sum_{i=1}^{N_p} g\left(\Theta_t^{(i)}\right) \overbrace{\frac{p\left(\Theta_t^{(i)}, \mathcal{D}_t\right)}{\pi\left(\Theta_t^{(i)}|\mathcal{D}_t\right)}}^{w_t^{(i)}} \left[\frac{1}{N_p} \sum_{i=1}^{N_p} \underbrace{\frac{p\left(\Theta_t^{(i)}, \mathcal{D}_t\right)}{\pi\left(\Theta_t^{(i)}|\mathcal{D}_t\right)}}_{w_t^{(i)}}\right]^{-1} \\
&= = \frac{\frac{1}{N_p} \sum_{i=1}^{N_p} g\left(\Theta_t^{(i)}\right) w_t^{(i)}}{\frac{1}{N_p} \sum_{i=1}^{N_p} w_t^{(i)}} = \sum_{i=1}^{N_p} g\left(\Theta_t^{(i)}\right) \overbrace{\frac{w_t^{(i)}}{\sum_{i=1}^{N_p} w_t^{(i)}}}^{\tilde{w}_t^{(i)}} \\
&= \sum_{i=1}^{N_p} \tilde{w}_t^{(i)} g\left(\Theta_t^{(i)}\right).
\end{aligned}$$

This result draws, as expected, parallels with the IS result in the previous section, and as such we establish the approximation of (6.7) as

$$p\left(\Theta_t|\mathcal{D}_t\right) \approx \hat{p}\left(\Theta_t|\mathcal{D}_t\right) = \sum \tilde{w}_t^{(i)} \delta\left(\Theta_t - \Theta_t^{(i)}\right), \tag{6.8}$$

with new *normalised importance weights*

$$\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{i=1}^{N_p} w_t^{(i)}} \qquad w_t^{(i)} = \frac{p\left(\Theta_t^{(i)}, \mathcal{D}_t\right)}{\pi\left(\Theta_t^{(i)}|\mathcal{D}_t\right)}, \tag{6.9}$$

and with the approximation set updated, as a new observation y_t arrives, as

$$\left\{\Theta_{t-1}^{(i)}, w_{t-1}^{(i)}\right\}_{i=1}^{N_p} \xrightarrow{y_t} \left\{\Theta_t^{(i)}, w_t^{(i)}\right\}.$$

The main problem with the above approach, is that for each new data point y_t, the entire trajectory Θ_{t-1} would need to be recalculated using IS. Also, even if it were possible to calculate the importance weights directly, this is not suitable for the requirements of sequential online estimation. As t evolves the computational burden would make it infeasible to approximate $p\left(\Theta_t|\mathcal{D}_t\right)$.

A solution is to decompose the proposal density $\pi\left(\Theta_t|\mathcal{D}_t\right)$ in a recursive way, such that

$$\pi\left(\Theta_t|\mathcal{D}_t\right) = \pi\left(\Theta_{t-1}|\mathcal{D}_{t-1}\right)\pi\left(\boldsymbol{\theta}_t|\Theta_{t-1},\mathcal{D}_t\right). \tag{6.10}$$

We can easily see that, by iteration, we have

$$\pi\left(\Theta_t|\mathcal{D}_t\right) = \pi\left(\boldsymbol{\theta}_0\right)\prod_{k=1}^{t}\pi\left(\boldsymbol{\theta}_t|\Theta_{k-1},\mathcal{D}_k\right).$$

By replacing the weights (6.9) with this factorisation we have

$$\begin{aligned} w_t^{(i)} &= \frac{p\left(\Theta_t^{(i)},\mathcal{D}_t\right)}{\pi\left(\Theta_t|\mathcal{D}_t\right)} \\ &= \frac{p\left(\Theta_t^{(i)},\mathcal{D}_t\right)}{\pi\left(\boldsymbol{\theta}_t^{(i)}|\Theta_{t-1}^{(i)},\mathcal{D}_t\right)\pi\left(\Theta_{t-1}^{(i)}|\mathcal{D}_{t-1}\right)}. \end{aligned}$$

Furthermore, using the Markovian factorisation of $p\left(\Theta_t,\mathcal{D}_t\right)$ in (3.8), we can finally write

the weights (6.9) in a recursive manner such that

$$\begin{aligned} w_t^{(i)} &= \frac{p\left(\Theta_t^{(i)}, \mathcal{D}_t\right)}{\pi\left(\boldsymbol{\theta}_t^{(i)}|\Theta_{t-1}^{(i)}, \mathcal{D}_t\right)\pi\left(\Theta_{t-1}^{(i)}|\mathcal{D}_{t-1}\right)} \\ &= \frac{p\left(\Theta_{t-1}^{(i)}, \mathcal{D}_{t-1}\right)p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right)p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}\right)}{\pi\left(\boldsymbol{\theta}_t^{(i)}|\Theta_{t-1}^{(i)}, \mathcal{D}_t\right)\pi\left(\Theta_{t-1}^{(i)}|\mathcal{D}_{t-1}\right)} \qquad (6.11) \\ &= \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right)p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}\right)}{\pi\left(\boldsymbol{\theta}_t^{(i)}|\Theta_{t-1}^{(i)}, \mathcal{D}_t\right)} \underbrace{\frac{p\left(\Theta_{t-1}^{(i)}, \mathcal{D}_{t-1}\right)}{\pi\left(\Theta_{t-1}^{(i)}|\mathcal{D}_{t-1}\right)}}_{w_{t-1}^{(i)}} \\ &= \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right)p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}\right)}{\pi\left(\boldsymbol{\theta}_t^{(i)}|\Theta_{t-1}^{(i)}, \mathcal{D}_t\right)} w_{t-1}^{(i)}. \qquad (6.12) \end{aligned}$$

This recursive formulation of the weights and importance density is crucial to enable online Bayesian filtering applying IS methods. We note, that although this method allows for an estimation of $p\left(\Theta_t|\mathcal{D}_t\right)$ as discussed in Chapter 3, it does not rely on the *predict-update* steps, but rather on a numerical approximation of the state posterior given the data.

Another important property of Bayesian filtering using IS methods follows from the fact (Pitt *et al.* (2012)) that we can also approximate the marginal likelihood

$$p\left(\mathcal{D}_t|\boldsymbol{\theta}_{0:t}\right) = p\left(y_1\right)\prod_{k=2}^{t} p\left(y_k|\mathcal{D}_{k-1}\right)$$

by using the unnormalised importance weights and calculating

$$\hat{L}_t = \hat{p}\left(\mathcal{D}_t|\boldsymbol{\theta}_{0:t}\right) = \prod_{k=1}^{t}\left(\frac{1}{N_p}\sum_{n=1}^{N_p} w_k^{(i)}\right).$$

It is shown in Crisan & Doucet (2002), however, that the variance will increase linearly over time.

This is a crucial result that will be used in Chapters 15 (SMC^2) and 13 (PMCMC). A direct application of this method to sequential state estimation is known as *Sequential Importance Sampling* (SIS, Kong *et al.* (1994)), which will be analysed in more detail, along with its inherent problems, in the next chapter.

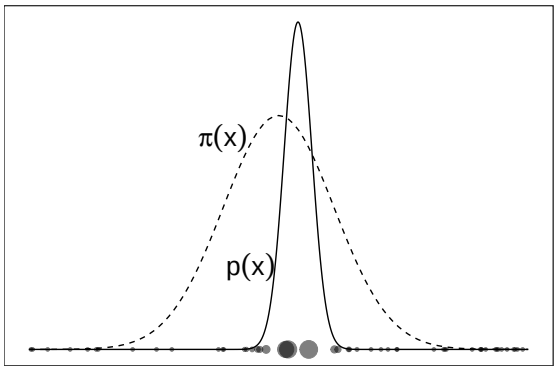


Figure 6.2: Illustration of Importance Sampling with $p(x)$ as the target and $\pi(x)$ as the importance distributions in the case where the target is highly peaked in comparison with the importance density. Circles indicate samples from $\pi(\cdot)$, with size proportional to the weight.

6.2 Proposals

The choice of proposal is critical for the performance of these methods. As an example, if the target is highly peaked in comparison to the importance density, a large proportion of the samples will have very low importance weight, as illustrated in Figure 6.2.

One of the advantages of SMC methods is that, in general, they provide a framework for state estimation which is independent of tasks such as the importance density design. This allows for the creation of a variety of algorithms, each tailored to the specific problem at hand. In the following sub-sections we will look at common proposal methods.

6.2.1 Optimal importance density

As we've seen in this section we are free to chose any importance density as long as it has a larger support than $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t)$ and it can factorised in a recursive fashion such as in (3.8). However, although the convergence of the IS algorithm will be guaranteed for $N_p \to \infty$ (Crisan & Doucet (2002)), in practice, ideally, we would want to choose an importance density which minimises the variance of the importance weights with a finite N_p. This importance density is commonly known as the *optimal importance density*. According to Doucet *et al.* (2001*b*), this density is

$$p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t) = \pi^\star(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, y_t), \tag{6.13}$$

which is shown in Doucet *et al.* (2001*b*) to minimise $\text{Var}\left[w_t^{(i)}|w_{t-1}^{(i)}, \boldsymbol{\theta}_{t-1}^{(i)}\right]$. To verify this, we can write the weights' variance as

$$\text{Var}\left[w_t^{(i)}|w_{t-1}^{(i)}, \boldsymbol{\theta}_{t-1}^{(i)}\right] = \int \left(w_t^{(i)}\right)^2 p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, y_t\right) d\boldsymbol{\theta}_t - \left(\int w_t^{(i)} p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, y_t\right) d\boldsymbol{\theta}_t\right)^2.$$

Using the relation in (6.12), we can rewrite as

$$\begin{aligned}
\text{Var}\left[w_t^{(i)}|w_{t-1}^{(i)}, \boldsymbol{\theta}_{t-1}^{(i)}\right] &= \int \left(\frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right) p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(i)}\right)}{p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}^{(i)}, y_t\right)} w_{t-1}^{(i)}\right)^2 p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, y_t\right) d\boldsymbol{\theta}_t \\
&\quad - \left(\int w_t^{(i)} p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, y_t\right) d\boldsymbol{\theta}_t\right)^2 \\
&= \int \left(w_{t-1}^{(i)}\right)^2 \frac{\left(p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right) p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(i)}\right)\right)^2}{p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}^{(i)}, y_t\right)} d\boldsymbol{\theta}_t \\
&\quad - \left(\int w_t^{(i)} p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, y_t\right) d\boldsymbol{\theta}_t\right)^2.
\end{aligned}$$

By substituting (6.13) in the above and simplifying, we get

$$\begin{aligned}
\text{Var}\left[w_t^{(i)}|w_{t-1}^{(i)}, \boldsymbol{\theta}_{t-1}^{(i)}\right] &= \left(w_{t-1}^{(i)}\right)^2 \left[\int \frac{\left(p\left(y_t|\boldsymbol{\theta}_t\right) p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(i)}\right)\right)^2}{\pi^\star\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(i)}, y_t\right)} d\boldsymbol{\theta}_t - \left(p\left(y_t|\boldsymbol{\theta}_{t-1}^{(i)}\right)\right)^2\right] \\
&= \left(w_{t-1}^{(i)}\right)^2 \left[p\left(y_t|\boldsymbol{\theta}_{t-1}^{(i)}\right) \underbrace{\int p\left(y_t, \boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(i)}\right) d\boldsymbol{\theta}_t}_{p\left(y_t|\boldsymbol{\theta}_{t-1}^{(i)}\right)} - \left(p\left(y_t|\boldsymbol{\theta}_{t-1}^{(i)}\right)\right)^2\right] \\
&= \left(w_{t-1}^{(i)}\right)^2 \left[\left(p\left(y_t|\boldsymbol{\theta}_{t-1}^{(i)}\right)\right)^2 - \left(p\left(y_t|\boldsymbol{\theta}_{t-1}^{(i)}\right)\right)^2\right] \\
&= 0.
\end{aligned}$$

If we take into account the recursive weight calculation of (6.12), and denote the optimal importance density as in (6.13), $\pi^\star\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathcal{D}_t\right)$, we see that

$$\begin{aligned}
w_t^{(i)} &= \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right) p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}\right)}{\pi^\star\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{0:t-1}^{(i)}, \mathcal{D}_t\right)} \\
&= \frac{p\left(\boldsymbol{\theta}_t^{(i)}, y_t|\boldsymbol{\theta}_{t-1}^{(i)}\right)}{p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}, y_t\right)} w_{t-1}^{(i)}.
\end{aligned}$$

Since $p\left(\boldsymbol{\theta}_t^{(i)}, y_t|\boldsymbol{\theta}_{t-1}^{(i)}\right)$ can be decomposed as

$$p\left(\boldsymbol{\theta}_t^{(i)}, y_t|\boldsymbol{\theta}_{t-1}^{(i)}\right) = p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}, y_t\right) p\left(y_t|\boldsymbol{\theta}_{t-1}^{(i)}\right),$$

the above weight can be further simplified into

$$\begin{aligned} w_t^{(i)} &= \frac{p\left(\boldsymbol{\theta}_t^{(i)}, y_t|\boldsymbol{\theta}_{t-1}^{(i)}\right)}{p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}, y_t\right)} w_{t-1}^{(i)} \\ &= \frac{p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}, y_t\right) p\left(y_t|\boldsymbol{\theta}_{t-1}^{(i)}\right)}{p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}, y_t\right)} w_{t-1}^{(i)} \\ &= p\left(y_t|\boldsymbol{\theta}_{t-1}^{(i)}\right) w_{t-1}^{(i)}. \end{aligned} \tag{6.14}$$

Since the weights at time t, $w_t^{(i)}$ are independent of $\boldsymbol{\theta}_t^{(i)}$, given $\boldsymbol{\theta}_{t-1}^{(i)}$,their variance will be

$$\text{Var}\left[w_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}, y_t\right] = 0.$$

However, it will be often the case where this density is not analytically available for sampling, except in a few cases. One such case is the NDLM, where the KF recursions can be employed for sampling for the importance density. In this case, the filter is said to be *fully adapted*.

If we consider a NDLM as formulated in Section 2.3.1 we will have the optimal importance density in the form

$$\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, y_t \sim \mathcal{N}\left(\bar{\boldsymbol{\theta}}_t, \mathsf{C}_t\right)$$

with

$$\begin{aligned} \bar{\boldsymbol{\theta}}_t &= \left(\mathbf{I} - \mathbf{K}_t\mathsf{F}^T\right)\boldsymbol{\theta}_{t-1} + \mathbf{K}_t y_t \\ \mathsf{C}_t &= \left(\mathbf{I} - \mathbf{K}_t\mathsf{F}^T\right)\mathsf{W} \end{aligned}$$

and

$$\mathbf{K}_t = \mathsf{W}\mathsf{F}^T\left(\mathsf{F}\mathsf{W}\mathsf{F}^T + \mathsf{V}\right)^{-1}.$$

That is, when using the optimal importance density, the weights will be proportional to

the *predictive likelihood* as described in (4.9). In the case of the NDLM, they will be

$$\begin{aligned} y_t|\boldsymbol{\theta}_{t-1} &\sim \mathcal{N}\left(y_t; f_t, Q_t\right) \\ f_t &= \mathsf{F}^T\mathsf{G}\boldsymbol{\theta}_{t-1} \\ Q_t &= \mathsf{F}^T\mathsf{W}\mathsf{F} + V \end{aligned}$$

A particle filter is said to be *exact* if i.i.d. samples can be drawn from the empirical filtering density, however as stated in Pitt & Shephard (1999), due to the discrete approximation to $p\left(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}\right)$, even fully adapted PFs might not produce i.i.d. samples from $p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right)$.

6.2.2 Prior

Although in general cases we cannot sample directly from (3.14), we can however create an alternative density from which it easy to sample, $\pi\left(\cdot\right)$, as in (6.3).

As we have seen in Section 6.2.1, the importance density should be chosen as to minimise $\text{Var}\left[\tilde{w}_t^{(i)}\right]$. According to Doucet *et al.* (2000) a common choice for $\pi\left(\cdot\right)$ is the prior itself, which in the DGLM case is given by (2.3):

$$\pi\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t\right) \approx p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\right) \overset{\text{DGLM}}{=} \mathcal{N}\left(\boldsymbol{\theta}_t; \mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}_t\right). \tag{6.15}$$

Using this importance density has two advantages. First, the fact that it is easy to sample from it. Second, using this importance density the importance weights calculation will simplify to being proportional to the likelihood. We can see from (6.12) that

$$\begin{aligned} w_t^{(i)} &\propto w_{t-1}^{(i)} \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right) p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}\right)}{\pi\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{0:t-1}^{(i)}, \mathcal{D}_t\right)} \\ &\propto w_{t-1}^{(i)} \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right) p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}\right)}{p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}\right)} \\ &\propto w_{t-1}^{(i)} p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right). \end{aligned}$$

Using the prior as an importance density presents some problems however. We are now not taking into account the current observation y_t when sampling. This is the reason why this proposal is called often a *blind proposal* in the literature. Sampling from the prior is one of the building blocks for one of the most well-known filtering algorithms, the *bootstrap filter*, which we will explore in more detail in Section 7.4 on page 91.

The convergence of this method is guaranteed by the central limit theorem and the error term is $\mathcal{O}\left(N^{-1/2}\right)$ regardless of the dimensionality of $\boldsymbol{\theta}$ (Liu (2002)).

6.2.3 Local Linearisation

An importance density for DGLMs can be obtained by local linearisation of the optimal importance density as detailed in Doucet *et al.* (2000). Doucet *et al.* (2000) starts by assuming that the optimal importance density is twice differentiable with regard to $\boldsymbol{\theta}_t$ and that we have the relation

$$\ell\left(\boldsymbol{\theta}_t\right) \triangleq \log p\left(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, y_t\right).$$

We start by defining the gradient and hessian, respectively, as

$$\nabla \ell\left(\boldsymbol{\theta}\right) = \left.\frac{\partial \ell\left(\boldsymbol{\theta}_t\right)}{\partial \boldsymbol{\theta}_t}\right|_{\boldsymbol{\theta}_t=\boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial \ell(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_{t,1}} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_{t,n}} \end{bmatrix} \tag{6.16}$$

$$\nabla^2 \ell\left(\boldsymbol{\theta}\right) = \left.\frac{\partial^2 \ell\left(\boldsymbol{\theta}_t\right)}{\partial \boldsymbol{\theta}_t \partial \boldsymbol{\theta}_t^T}\right|_{\boldsymbol{\theta}_t=\boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial^2 \ell(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_{t,1}^2} & \cdots & \frac{\partial^2 \ell(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_{t,1} \partial \boldsymbol{\theta}_{t,n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_{t,n} \partial \boldsymbol{\theta}_{t,1}} & \cdots & \frac{\partial^2 \ell(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_{t,n}^2} \end{bmatrix}. \tag{6.17}$$

Recalling that for a second order Taylor expansion, we have

$$f\left(\mathbf{x} - \delta\mathbf{x}\right) = f\left(\mathbf{x}\right) + \nabla f\left(\mathbf{x}\right)^T \delta\mathbf{x} + \frac{1}{2}\delta\mathbf{x}^T \nabla^2 f\left(\mathbf{x}\right) \delta\mathbf{x} + \ldots$$

it follows that for $\ell\left(\boldsymbol{\theta_t}\right)$, performing the expansion in $\boldsymbol{\theta}$, we will have

$$\ell\left(\boldsymbol{\theta}_t\right) \approx \ell\left(\boldsymbol{\theta}\right) + \nabla \ell\left(\boldsymbol{\theta}\right)^T \left(\boldsymbol{\theta}_t - \boldsymbol{\theta}\right) + \frac{1}{2}\left(\boldsymbol{\theta}_t - \boldsymbol{\theta}\right)^T \nabla^2 \ell\left(\boldsymbol{\theta}\right)\left(\boldsymbol{\theta}_t - \boldsymbol{\theta}\right).$$

Doucet *et al.* (2000) notes that by setting

$$\begin{aligned} \Sigma\left(\boldsymbol{\theta}\right) &= -\nabla^2 \ell\left(\boldsymbol{\theta}\right)^{-1} \\ \mathbf{m}\left(\boldsymbol{\theta}\right) &= \Sigma\left(\boldsymbol{\theta}\right) \nabla \ell\left(\boldsymbol{\theta}\right), \end{aligned}$$

we get the following result

$$\ell\left(\boldsymbol{\theta}_t\right) \approx K - \frac{1}{2}\left(\boldsymbol{\theta}_t - \boldsymbol{\theta} - \mathbf{m}\left(\boldsymbol{\theta}\right)\right)^T \Sigma^{-1}\left(\boldsymbol{\theta}\right)\left(\boldsymbol{\theta}_t - \boldsymbol{\theta} - \mathbf{m}\left(\boldsymbol{\theta}\right)\right).$$

According to Doucet *et al.* (2000) this leads to the locally linearised importance density as

$$\pi\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, y_t\right) = \mathcal{N}\left(\boldsymbol{\theta}_t; \mathbf{m}\left(\boldsymbol{\theta}\right) + \boldsymbol{\theta}, \Sigma\left(\boldsymbol{\theta}\right)\right).$$

For the specific case of the DGLM, we use the canonical form in (2.5)

$$y_t \sim p\left(y_t|\eta_t\right) = \exp\left\{\frac{z\left(y_t\right)\eta_t - b\left(\eta_t\right)}{a\left(\phi_t\right)} + c\left(y_t, \phi_t\right)\right\}$$
$$\boldsymbol{\theta}_t = \mathsf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim \mathcal{N}\left(\mathbf{0}, \mathsf{W}\right)$$

and the log-likelihood can then be written as

$$\begin{aligned}
\ell\left(\boldsymbol{\theta_t}\right) &= \log p\left(y_t|\boldsymbol{\theta}_t\right) + \underbrace{\log p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\right)}_{\text{Normal state transition model}} + K \\
&= \underbrace{a^{-1}\left(\phi_t\right) z\left(y_t\right) \mathsf{F}^T\boldsymbol{\theta}_t - b\left(\mathsf{F}^T\boldsymbol{\theta}_t\right) + c\left(y_t, \phi_t\right)}_{\log p(y_t|\eta_t)} \\
&\quad \underbrace{-\frac{1}{2}\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right)^T \mathsf{W}^{-1}\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right)}_{\log p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})} + K.
\end{aligned}$$

We can now calculate (6.16) and (6.17):

$$\begin{aligned}
\nabla\ell\left(\boldsymbol{\theta}\right) &= \left.\frac{\partial\ell\left(\boldsymbol{\theta}_t\right)}{\partial\boldsymbol{\theta}_t}\right|_{\boldsymbol{\theta}_t=\boldsymbol{\theta}} \\
&= \left.\frac{\partial\left(a^{-1}\left(\phi_t\right) z\left(y_t\right)\mathsf{F}^T\boldsymbol{\theta}_t - b\left(\mathsf{F}^T\boldsymbol{\theta}_t\right) + c\left(y_t, \phi_t\right)\right)}{\partial\boldsymbol{\theta}_t}\right|_{\boldsymbol{\theta}_t=\boldsymbol{\theta}} \\
&\quad \left. - \frac{1}{2}\frac{\partial\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right)^T \mathsf{W}^{-1}\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right)}{\partial\boldsymbol{\theta}_t}\right|_{\boldsymbol{\theta}_t=\boldsymbol{\theta}} \\
&= z\left(y_t\right)\mathsf{F}^T - \nabla b\left(\mathsf{F}^T\boldsymbol{\theta}_t\right) - \mathsf{W}^{-1}\left(\boldsymbol{\theta} - \mathsf{G}\boldsymbol{\theta}_t\right)
\end{aligned}$$

and

$$\begin{aligned}
\nabla^2\ell\left(\boldsymbol{\theta}\right) &= \left.\frac{\partial^2\ell\left(\boldsymbol{\theta}_t\right)}{\partial\boldsymbol{\theta}_t\partial\boldsymbol{\theta}_t^T}\right|_{\boldsymbol{\theta}_t=\boldsymbol{\theta}} \\
&= \left.\frac{\partial^2 b\left(\eta_t\right)}{\partial\boldsymbol{\theta}_t\partial\boldsymbol{\theta}_t^T}\right|_{\boldsymbol{\theta}_t=\boldsymbol{\theta}} - \mathsf{W}^{-1} \\
&= -\nabla^2 b\left(\eta_t\right) - \mathsf{W}^{-1}.
\end{aligned}$$

Since the form of $b\left(\cdot\right)$ is well known from Section 2.3, the gradient and hessian can be calculated.

6.2.4 Conjugate filtering

Another possibility in terms of proposal densities that might be useful especially in the case of non-linear DGLMs is to use the moment approximations of the conjugate filter as detailed in Chapter 5.

As we have seen in Algorithm 5.1 on page 52, given an approximation of the moments at time $t-1$, such that

$$\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1} \sim [\mathbf{m}_{t-1}, \mathsf{C}_{t-1}],$$

it is possible to apply the CF recursions to approximate the moments at time t, that is

$$\boldsymbol{\theta}_t|\mathcal{D}_t \sim [\mathbf{m}_t, \mathsf{C}_t],$$

where the moments can be updated (as described in detail in Section 5) by

$$\begin{aligned}\mathbf{m_t} &= \mathbf{a}_t + \mathsf{R}_t\mathsf{F}\frac{(f_t^\star - f_t)}{q_t} \\ \mathsf{C}_t &= \mathsf{R}_t - \mathsf{R}_t\mathsf{F}\left(1 - \frac{q_{\star t}}{q_t}\right)\frac{1}{q_t}\mathsf{F}^T\mathsf{R}_t.\end{aligned}$$

By using a multivariate normal proposal with the approximated moments at time t, such that

$$\pi\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t\right) \approx \mathcal{N}\left(\mathbf{m}_t, \mathsf{C}_t|\mathcal{D}_t\right)$$

any discrepancy introduced by the fact the we are using a linear approximation to a non-linear importance density can be corrected by the importance sampling weights. Namely, given that the importance weights are calculated according to (6.12), that is

$$w_t^{(i)} = \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right)p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}\right)}{\pi\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{0:t-1}^{(i)}, \mathcal{D}_t\right)}w_{t-1}^{(i)},$$

we can now simply replace with the known values for the DGLM case, that is

$$w_t^{(i)} = \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right)\mathcal{N}\left(\boldsymbol{\theta}_t^{(i)}|\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}\right)}{\mathcal{N}\left(\boldsymbol{\theta}_t^{(i)}|\mathbf{m}_t, \mathsf{C}_t\right)}w_{t-1}^{(i)}. \tag{6.18}$$

The likelihood, $p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right)$ and quantities $q_t, q_t^\star, f_t, f_t^\star$ will clearly depend on the class of linear DGLM chosen as detailed in Chapter (5). In Section 7.4 on page 91 we will look at examples of particle filters making use of this specific proposal distribution.

6.2.5 Extended Kalman Filter

As discussed in Section 4.2, the adjusted model of Fahrmeir (1992) allows us to approximate the non-linear DGLM to a linear adjusted model. This allows the approximation of the filtering density with a Gaussian density such that

$$p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right) \approx \mathcal{N}\left(\mathbf{m}_t, \mathsf{C}_t\right).$$

As in the previous section, this approximation can be used to construct a proposal density such that

$$\pi\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t\right) \approx \mathcal{N}\left(\mathbf{m}_t, \mathsf{C}_t|\mathcal{D}_t\right).$$

The recursions of 4.2 (namely in Algorithm 4.3) can be used to update the moments (conditioned on $\boldsymbol{\theta}_{t-1}$) and the importance weights calculated using

$$w_t^{(i)} = \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right)\mathcal{N}\left(\boldsymbol{\theta}_t^{(i)}|\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}\right)}{\mathcal{N}\left(\boldsymbol{\theta}_t^{(i)}|\mathbf{m}_t, \mathsf{C}_t\right)} w_{t-1}^{(i)}. \tag{6.19}$$

One of the potential advantages of this proposal density (such as the proposal in Section 6.2.4) is the fact that the observation is included in the proposal construction.

Although, for computational reasons, we have not considered an iterative process to approximate the mode of the linearised density, any potential mismatch would in theory be corrected by the importance weights in (6.19). With the EKF and CF proposals in (6.19) and (6.18) the defined transition density function was the Normal distribution. An alternative would be to use a multivariate Student-t (MVt) with ν degrees of freedom since, for instance, the heavier tails could potentially alleviate the Normal distribution's sensitivity to outliers. The MVt degrees of freedom would need, however, to be carefully chosen. Using a low value (such as a Cauchy distribution, in the $\nu = 1$ case) might overdisperse the particles and move them to regions far from the true posterior mean, especially in high dimensional models (Cappé *et al.* (2006)). Given that both of these proposals (CF and EKF) try to perform *adaptation*, by incorporating the current observation in the state update, we have used the Normal version of the proposal density in the examples and case studies.

6.2.6 Other proposals

The study of importance densities for IS in SSMs is an active area with a wealth of choice of methods each with its advantages for a specific domain. Proposals based on the Unscented Kalman Filter (UKF, Julier & Uhlmann (1997)) which results suggest provides

a better approximation than the EKF in the presence of strong non-linearities. This forms the basis of the Unscented Particle Filter (UPF, Van Der Merwe *et al.* (2001)). Other importance densities include Laplace approximations (Moulines (2004)) and the Split-Gaussian proposals (Kokkala & Sarkka (2015)) among others.

Chapter 7

Sequential Monte Carlo

7.1 Sequential Importance Sampling

The simplest implementation of the IS methods described in Section 6.1 to perform state estimation is the Sequential Importance Sampling (SIS) algorithm (presented in Algorithm 7.1) introduced in Kong *et al.* (1994). SIS is a straight-forward implementation of the IS method for Bayesian filtering in Section 6.1, aiming at building an approximation of $p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right)$ sequentially, starting from a state prior in the form of (2.4).

In the specific instance of DGLMs, an appropriate implementation of the importance density can be chosen. For instance, in the case of using the prior as $\pi\left(\cdot\right)$, Algorithm 7.1 can be implemented by adding the intermediate step of calculating

$$\eta_t^{(i)} = g^{-1}\left(\lambda_t^{(i)}\right)$$

with $\lambda_t^{(i)} = F^T\boldsymbol{\theta}_t^{(i)}$ and using the appropriate likelihood, $p\left(y_t|\eta_t^{(i)}\right)$ according the DGLM instance (Poisson, Binomial, *etc.*) for calculation of the unnormalised weights $w_t^{(i)}$.

Example. SIS applied to a $\mathcal{M} = \{\mathcal{P}\left(1\right)\}$ NDLM.

We consider a $\mathcal{M} = \{\mathcal{P}\left(1\right)\}$ NDLM with $\tau^2 = 0.3$ and $\nu^2 = 4.3$ and apply the SIS method with $N_p = 500$. The importance density chosen was the prior itself, as described in Section 6.2.2, that is $\theta_t|\theta_{t-1}^{(i)} \sim \mathcal{N}\left(\theta_{t-1}^{(i)}, \tau^2\right)$. The time evolution of the weights represented in Figure 7.1a on page 84 and the individual state particle trajectories are represented in Figure 7.1b on page 84.

Algorithm 7.1 Sequential Importance Sampling (SIS)

initialisation Given m_0, C_0, $\boldsymbol{\theta}_0^{(i)} \sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0)$ and $\left\{\tilde{w}_0^{(i)}\right\}_{i=1}^{N_p} = \frac{1}{N_p}$.

for $t \leftarrow 1$ to N_{obs}

for $i \leftarrow 1$ to N_p

sample $\boldsymbol{\theta}_t^{(i)} \sim \pi\left(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t\right)$

calculate unnormalised importance weights

$$w_t^{(i)} = w_{t-1}^{(i)} \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right) p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}\right)}{\pi\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{0:t-1}^{(i)}, \mathcal{D}_t\right)}$$

normalise the importance weights

$$\tilde{w}_t = \frac{w_t^{(i)}}{\sum_{j=1}^{N_p} w_t^{(i)}}$$

compute state estimate

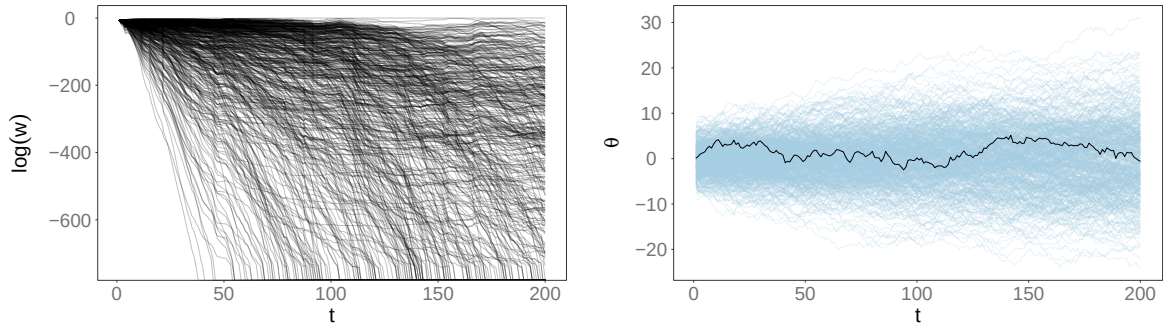
$$\hat{p}\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right) = \sum_{i=1}^{N_p} \tilde{w}_t^{(i)} \delta\left(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(i)}\right)$$

7.1.1 Weight degeneracy

One of the main problems with IS methods, particularly evident in the SIS filter, which makes it unsuitable for most real world applications, is that after a few iterations almost all weights will tend to zero. This is commonly known in the literature as *weight degeneracy*. As we can see from Figure 7.1a on the next page, although all the weights have at $t = 0$ a uniform value of $\left\{\tilde{w}_0^{(i)}\right\}_{i=1}^{N_p} = 1/N_p$, most of the weights quickly become negligible, indicating that only a few particles are effectively contributing to the posterior estimate of $p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right)$. This is due to the fact that as t evolves the variance of the weights will increase and will eventually lead to a collapse of the marginal posterior $p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right)$. If we consider and importance density of the form of (6.10), the increase of the weight variance is inevitable (Kong *et al.* (1994)).

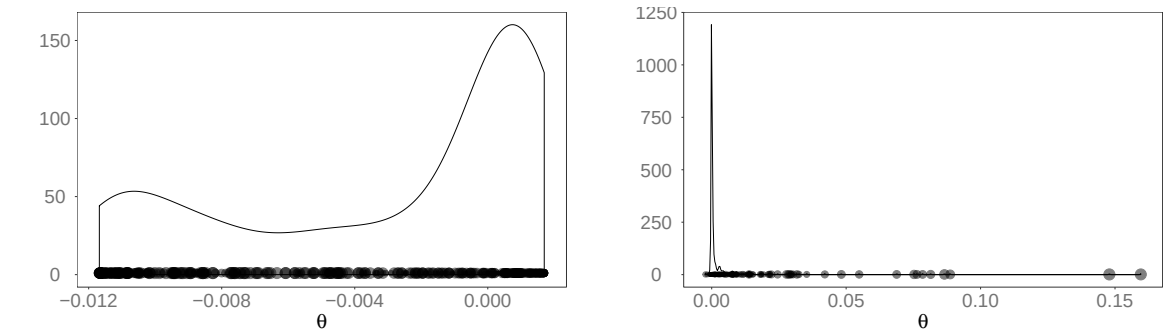
We can see this effect by taking into account the example model on the facing page and comparing the marginals at time $t = 2$ and $t = 10$ in Figures 7.2a and 7.2b.

Although one of the strategies to mitigate this degeneracy, the optimal choice of importance density, was covered in Section 6.2, other methods are available, such as particle resampling, which will be discussed in Section 7.3. Another possibility to reduce the



(a) $\tilde{w}_{0:t}^{(i)}$ (b) $\theta_{0:t}^{(i)}$

Figure 7.1: Importance weights (*left*, log scale) and state for individual particles (*right*, realisation's "true" state in black), $\left\{\theta_t^{(i)}, w_t^{(i)}\right\}_{i=1}^{N_p}$ for $t = 1, \ldots, N_{obs}$ from a SIS filter for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with $N_{obs} = 200$ and $N_p = 500$ using $p(\theta|\theta_{t-1}, \boldsymbol{\Phi})$ as the importance density.



(a) $\hat{p}(\theta_2|\mathcal{D}_2)$ (b) $\hat{p}(\theta_{10}|\mathcal{D}_{10})$

Figure 7.2: States for individual particles, $\left\{\theta_t^{(i)}\right\}_{i=1}^{N_p}$ at $t = 2$ (*left*) and $t = 10$ (*right*) from a SIS filter for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with $N_{obs} = 200$ and $N_p = 500$ using $p(\theta|\theta_{t-1}, \boldsymbol{\Phi})$ as the importance density.

weights' degeneracy relies on "brute force". Since we've seen in Section 6 the approximation will tend to the true posterior as $N_p \to \infty$, then, an increase in the number of particles will temporarily delay the degeneracy problem. However, this is an obvious issue in computational terms, especially when considering real-time estimations. Although (for the same model and data), SMC methods will usually display a $\mathcal{O}(N_p)$ behaviour[1], a compromise must be achieved between the number of particles and time constraints (as we will examine in Section 16.1).

[1] In a naïve implementation, not considering factors such as parallelisation.

7.2 Effective Sample Size

A common way of quantifying weight degeneracy, as described in Kong *et al.* (1994); Liu (1996*a*), is to estimate the *Effective Sample Size* (ESS) at each time step.

If we consider the proposal, $\pi(\cdot)$ and the target distribution $p(\cdot)$, and $\bar{g}_{IS}$ and $\bar{g}_{MC}$ as approximations of (6.1) but using IS and sampling from the true posterior, respectively, we can define an *efficiency ratio* as

$$\eta_t(\pi, p) = \frac{\mathrm{Var}\left[\bar{g}_{IS}|\mathcal{D}_t\right]}{\mathrm{Var}\left[\bar{g}_{MC}|\mathcal{D}_t\right]}.$$

According to Liu & Chen (1998); Kong *et al.* (1994); Liu (1996*a*), the ESS can then be written as

$$ESS = \frac{N_p}{\eta_t(\pi, p)}$$

with the desired property that when degeneracy increases, the variance of the IS estimate, $\mathrm{Var}\left[\bar{g}_{IS}|\mathcal{D}_t\right]$, will also increase, resulting in a lower ESS. One of the main problems with this definition is that the ESS cannot be calculated directly and as such must be estimated. However, according[2] to Liu (1996*b*) we can approximate

$$\frac{\mathrm{Var}\left[\bar{g}_{IS}|\mathcal{D}_t\right]}{\mathrm{Var}\left[g|\mathcal{D}_t\right]} \approx \frac{\mathrm{Var}\left[w_t^\star|\mathcal{D}_t\right] + 1}{N_p},$$

where $w^\star$ are the "true" importance weights, which are still not available for direct calculation. Nonetheless, the ESS can now be written as

$$\begin{aligned} ESS &= \frac{N_p}{\eta_t(\pi, p)} \\ &= N_p \left[\frac{\mathrm{Var}\left[\bar{g}_{IS}|\mathcal{D}_t\right]}{\mathrm{Var}\left[\bar{g}_{MC}|\mathcal{D}_t\right]}\right]^{-1} \\ &= N_p \left[\frac{\mathrm{Var}\left[\bar{g}_{IS}|\mathcal{D}_t\right]}{\mathrm{Var}\left[g|\mathcal{D}_t\right]}\right]^{-1} \\ &\approx \left[\frac{\mathrm{Var}\left[w_t^\star|\mathcal{D}_t\right] + 1}{N_p}\right]^{-1} \\ &= \frac{N_p}{\mathrm{Var}\left[w_t^\star|\mathcal{D}_t\right] + 1}. \end{aligned}$$

If we take the additional approximation that

$$\mathrm{Var}_\pi\left[w_t^\star|\mathcal{D}_t\right] + 1 \approx \mathrm{E}_\pi\left[w_t^{\star 2}|\mathcal{D}_t\right]$$

[2] For a full derivation, *cf.* Liu (1996*b*).

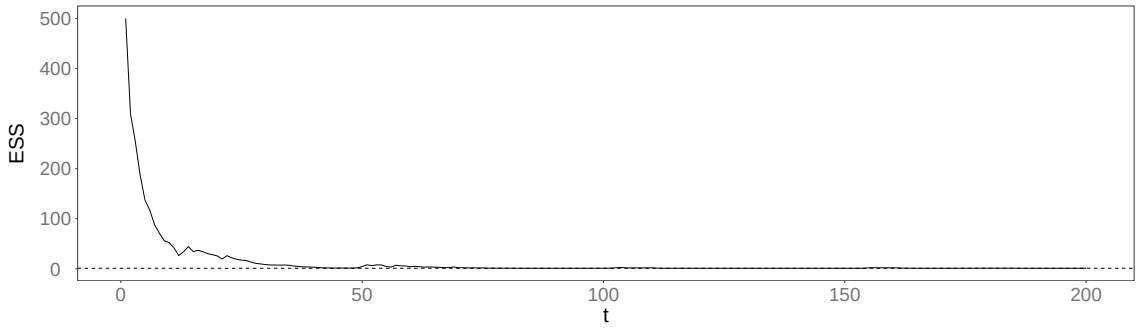


Figure 7.3: $\widehat{ESS}$ for SIS in a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with $N_{obs} = 200$ and $N_p = 500$.

and, by establishing a parallel between (6.2) and (6.4),

$$N_p \tilde{w}_t^{(i)} \approx w_t^{\star(i)},$$

we can then write an estimate for the ESS as

$$\begin{aligned}
\widehat{ESS} &= \frac{N_p}{\mathrm{Var}\left[w_t^{\star}|\mathcal{D}_t\right] + 1} \\
&= \frac{N_p}{\mathrm{E}_\pi\left[w_t^{\star 2}|\mathcal{D}_t\right]} \\
&= \frac{N_p}{\frac{1}{N_p}\sum_{i=1}^{N_p}\left(N_p \tilde{w}_t^{(i)}\right)^2} \\
&= \frac{1}{\sum_{i=1}^{N_p}\left(\tilde{w}_t^{(i)}\right)^2}.
\end{aligned} \tag{7.1}$$

From the ESS expression in (7.1) it is clear that it will take values between the extremes $\widehat{ESS} = N_p$, which can be interpreted as all N_p particles contributing equally to the density estimation, and $\widehat{ESS} = 1$, interpreted as a single particle contributing to the density estimation. Using this definition of ESS, we can then express it as a ratio of the total number of particles (*e.g.* half the particles contributing to the estimation would be equivalent to $\widehat{ESS} = N_p/2$).

Applying the ESS calculation to the SIS estimation in the example on page 82, we get the result in Figure 7.3.

From Figure 7.3, we can clearly see that the ESS starts from a value of N_p, as we have all particles initially with an uniform weight of $1/N_p$, decaying (exponentially) to the value of $\widehat{ESS} = 1$, interpreted as a single particle contributing to the estimation. Another related quantity is the *Effective Factor* (EF) measure which is the ESS normalised with

regard to the total number of particles, that is

$$\widehat{EF} = \frac{\widehat{ESS}}{N_p}.$$

7.3 Resampling methods

One of the main methods to deal with the problem of weight degeneracy is *resampling*, first introduced in the context of SMC in the seminal Gordon *et al.* (1993) paper.

As the name *resampling* suggests, the IS sampled particle set will be subject to another sampling stage, this time where particles will be selected proportionally to his weights. This step aims at selecting particles in such a way that minimises the weight variation by keeping particles concentrated around regions of higher likelihood. One of the consequences, in terms of expressing the state approximation, is that this allows us to remove the explicit dependency of the marginal state IS approximation of $p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right)$ in (6.8) from a weighted to an unweighted approximation (Hol *et al.* (2006)) such that

$$\sum_{i=1}^{N_p} \tilde{w}_t^{(i)} \delta\left(\boldsymbol{\theta}_{0:t} - \boldsymbol{\theta}_{0:t}^{(i)}\right) \longrightarrow \frac{1}{N_p} \sum_{k=1}^{N_p} D^{(k)} \delta\left(\boldsymbol{\theta}_{0:t} - \boldsymbol{\theta}_{0:t}^{(k)}\right), \tag{7.2}$$

where $D^{(k)}$ is the number of offspring of particle $\boldsymbol{\theta}_{0:t}^{(k)}$ after resampling. Intuitively, it can be noted that a particle k with no replications ($D^{(k)} = 0$) will be removed from the approximation. If we consider the IS approximation, in the form of (6.7), with and without resampling respectively as $\bar{g}_{ISR}$ and $\bar{g}_{IS}$, Douc *et al.* (2005) proves that the asymptotic convergence properties of the approximation when $N_p \to \infty$ are maintained when

$$\mathrm{E}\left[\left(\bar{g}_{IS} - \bar{g}_{ISR}\right)^2\right] \overset{N_p \to \infty}{\longrightarrow} 0.$$

Regarding the implementation of the resampling mechanism, although this an area of extensive research with many algorithms available, we will focus on the most common in the literature, namely Multinomial, Systematic and Stratified resampling. These three methods share the following characteristics:

- *Unbiasedness*
  Considering the definition in (7.2) unbiasedness means that despite resampling, the mean of the $p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right)$ is maintained (Douc *et al.* (2005)), such that

$$\mathrm{E}\left[D^{(i)}\right] = N_p w_t^{(i)}. \tag{7.3}$$

That is, after resampling, the duplication of $\boldsymbol{\theta}_t^{(i)}$ is expected to be $N_p w_t^{(i)}$. This particular characteristic will play a crucial role in guaranteeing the unbiasedness of the state estimation itself which is one of the fundamental criteria in methods analysed in subsequent sections.

- *Linear time*
  The considered algorithms will have a computational cost proportional to the number of particles N_p, that is they will operate in $\mathcal{O}(N_p)$.
- *Constant particle number*
  All of the considered resampling algorithms maintain the number of particles after resampling. The discrete approximation of $p(\boldsymbol{\theta}_t|\mathcal{D}_t)$ will be obtained from the same number of samples, N_p, as prior to resampling.

Resampling can be sequentially applied, at each time step t, and will form the basis of the Sequential Importance Resampling (SIR) algorithm, which will be detailed in Section 7.4. An important consequence of the resampling step, as we can see from (7.2), is that the new, resampled particles will now have uniform weights of $\left\{\tilde{w}_t^{(i)}\right\}_{i=1}^{N_p} = 1/N_p$. Given the discrete approximation set at time t as $\left\{\boldsymbol{\theta}_t^{(i)}, \tilde{w}_t^{(i)}\right\}_{i=1}^{N_p}$ we can then define a generic resampling distribution $\mathcal{R}(\cdot)$ such that the new resampled particle indices will be

$$k \sim \mathcal{R}\left(i = \{1, \ldots, N_p\}, \tilde{w}_t^{(i)} = \left\{\tilde{w}_t^{(1)}, \ldots, \tilde{w}_t^{(N_p)}\right\}\right). \tag{7.4}$$

The new discrete approximation set will then be

$$\left\{\boldsymbol{\theta}_t^{(k)}, \frac{1}{N_p}\right\}_{k=1}^{N_p}.$$

An important issue is when to perform resampling. Two strategies are available, as described in Doucet *et al.* (2001*a*), namely *static* and *dynamic checkpoints*.

For *static checkpoints*, we perform resampling at every $n > 0$ time steps whereas for *dynamic checkpoints*, we calculate the $\widehat{ESS}$, as defined in Section 7.2, at every step, as a measure of degeneracy and if it is below a certain pre-defined threshold, which we will call N_{eff} then we perform resampling. Although the choice of N_{eff} will left open and dependent on the case at hand, a common value is $\widehat{ESS} \leq N_{eff} = N_p/2$.

Although numerous resampling algorithms (including parallel implementations, such as in Murray *et al.* (2016)) are available in the literature and this constitutes an active research field, we will focus in this thesis, and the following sections, on the most prevalent, all within the constraints stated above.

7.3.1 Multinomial resampling

Multinomial resampling is possibly the most common method employed in the literature. This method samples from the categorical distribution $\mathcal{C} = \{1, 2, \ldots, N_p\}$ where the probability of selecting a particle i is $p(k = i) = \tilde{w}^{(k)}$. We can then iteratively generate a random uniform number

$$u_k \sim \mathcal{U}(0, 1),$$

and defining the cumulative sum of the weights as

$$Q_i = \sum_{s=1}^{i} w_t^{(s)}$$

we assign k as

$$k = \inf\{i : Q_i > u_k\}$$

Multinomial resampling is not the most efficient resampling algorithm (as shown in Carpenter *et al.* (1999)) with a computational complexity of $\mathcal{O}(N_p \log N_p)$. However, as demonstrated in Carpenter *et al.* (1999), it is possible to implement it in $\mathcal{O}(N_p)$ operations. This implementation, usually called the *inverse CDF* method (Douc *et al.* (2005)), consists in generating N_p ordered random numbers as

$$\begin{aligned} u_k &= u_{k+1} \sqrt[k]{\tilde{u}_k}, \qquad \tilde{u}_k \sim \mathcal{U}[0, 1) \\ u_{N_p} &= \sqrt[N_p]{\tilde{u}_{N_p}}, \end{aligned}$$

and calculate the new indices based on the generalised inverse of the CDF, that is, $F^{-1}(u_k)$. Multinomial resampling corresponds naturally with the Multinomial distribution where we have the new particle indices k sampled from

$$k \sim \mathcal{M}\left(\tilde{w}_t^{(1)}, \tilde{w}_t^{(2)}, \ldots, \tilde{w}_t^{(N_p)}\right), \qquad \sum_{i=1}^{N_p} \tilde{w}_t^{(i)} = 1.$$

7.3.2 Systematic resampling

Another resampling method, known as *Systematic resampling*, as presented in Carpenter *et al.* (1999) builds on the Multinomial resampling procedure in the previous section. In this case however, instead of sampling directly $u_k \sim \mathcal{U}(0, 1)$ a single random uniform is drawn $u \sim \mathcal{U}(0, 1)$ and the intermediate values

$$u_k = \frac{(k - 1) + u}{N_p}$$

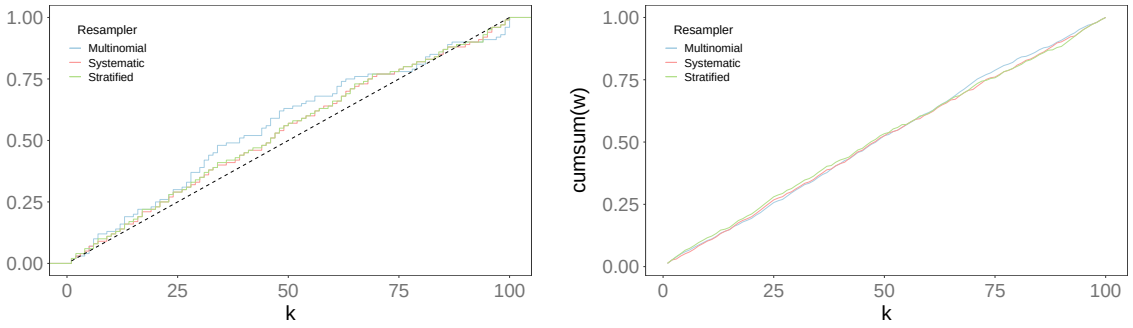
are constructed. In effect this is equivalent to partitioning the the random variable u domain, $(0, 1)$, into k equally spaced partitions $\left\{u + \frac{k}{N_p}\right\}_{k=1}^{N_p}$ and distributing the ordered weights according to these partitions with the number of weights per partition being counted. The index assignment is then the same as in the Multinomial case. According to Carpenter *et al.* (1999), Systematic resampling runs in $\mathcal{O}(N_p)$ time.

7.3.3 Stratified resampling

Stratified resampling, as presented in Press & Farrar (2012), bears similarities with Systematic resampling. In this case, the intermediate quantity u_k is calculated by drawing from a random uniform for each k such that

$$u_k = \frac{(k-1) + \bar{u}_k}{N_p}, \qquad \bar{u}_k \sim \mathcal{U}(0, 1).$$

We can see the similarities with Systematic resampling in partitioning the u domain and as in the Systematic method, the index assignment is then the same as in the Multinomial case. The objective of the uneven partitioning is to achieve a more uniform weight distribution across partitions. In Douc *et al.* (2005) it is shown that the conditional variance of both systematic and stratified resampling is, in theory, always smaller than that of Multinomial resampling.



(a) (b)

Figure 7.4: Indices k cumulative distribution function (*left*) from resampling $n = 100$ samples from $\mathcal{U}(0, 1)$ with different resamplers (dashed line is identity line) and resulting cumulative sum of normalised weights (*right*)

Regarding the mean computational cost of these methods, we can see in Figure 7.5 the linear $O(N_p)$ behaviour for 500 runs on resampling N items with weights distributed from $\mathcal{U}(0, 1)$. If we consider $N = 100$ random weights sampled from a uniform distribution, such that

$$w_k \sim \mathcal{U}(0, 1), \qquad i = 1, \cdots, N,$$

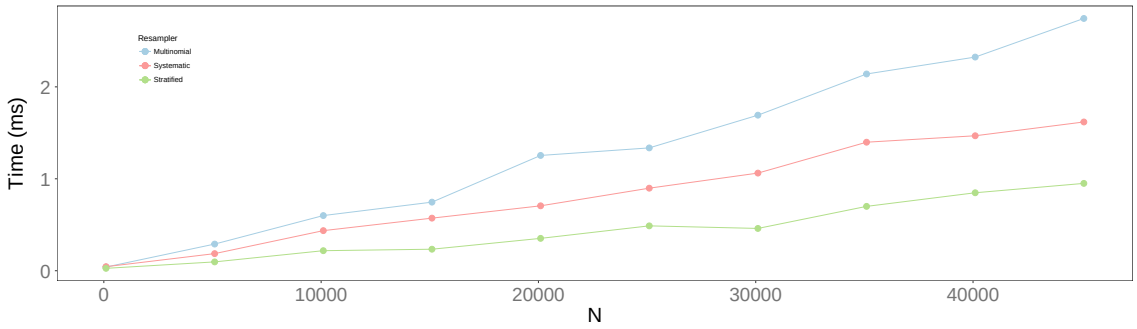


Figure 7.5: Mean computational time for a resample using Multinomial, Systematic and Stratified resampling for N_p samples from $\mathcal{U}(0,1)$ (starting at $N_p = 10$).

and normalise the weights according to

$$\tilde{w}_k = \frac{w_k}{\sum_{i=1}^{N} w_k}, \qquad i = 1, \ldots, N.$$

Then, by applying the various resampling methods, we would expect to see a uniform distribution of selected indices, according to (7.4). This result is presented in Figure 7.4.

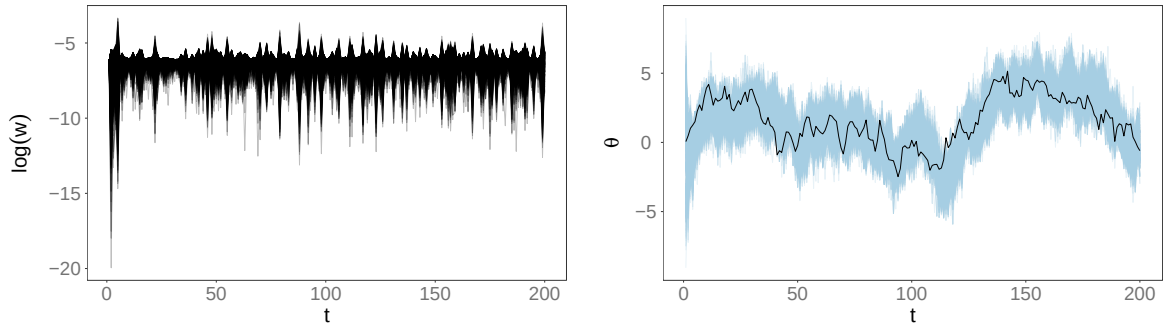
These stratified and systematic results are in accordance with Douc *et al.* (2005), which states that stratified and systematic resampling methods yield comparable results. In the "real world" case studies in Part V we will however use mainly stratified resampling due to the combination of lower computational cost (compared to multinomial) and the fact that a central limit theorem has been established.

7.4 Sequential Importance Resampling

Sequential Importance Resampling (SIR) methods, introduced by Gordon *et al.* (1993) and Kitagawa (1996), employ resampling methods as described in Section 7.3, to try to delay the degeneracy problem.

The SIR algorithm (detailed in Algorithm 7.2) aims at calculating a discrete approximation of the state posterior, provided with known model parameters. An estimate of the state, along with the weight evolution, for the same model as used for example on page 82 is displayed in Figure 7.6. In this example, the resample step was applied using a static threshold with $n = 1$ (that is, resampling at every step t) with a Multinomial resampler and using the prior, $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ as the importance density. The difference between the SIS and SIR in terms of weight degeneracy is clear when we compare the ESS estimate for both filters represented in Figures 7.3 on page 86 and 7.7 on the following page.

This particular implementation (*sample-resample*, with the prior as the importance



(a) $w_{0:t}^{(i)}$ (b) $\theta_{0:t}^{(i)}$

Figure 7.6: Weights (*log scale*) and states for individual particles, $\left\{\theta_t^{(i)}, w_t^{(i)}\right\}_{i=1}^{N_p}$ for $t = 1, \ldots, N_{obs}$ from a SIR filter for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with $N_{obs} = 200$ and $N_p = 500$.

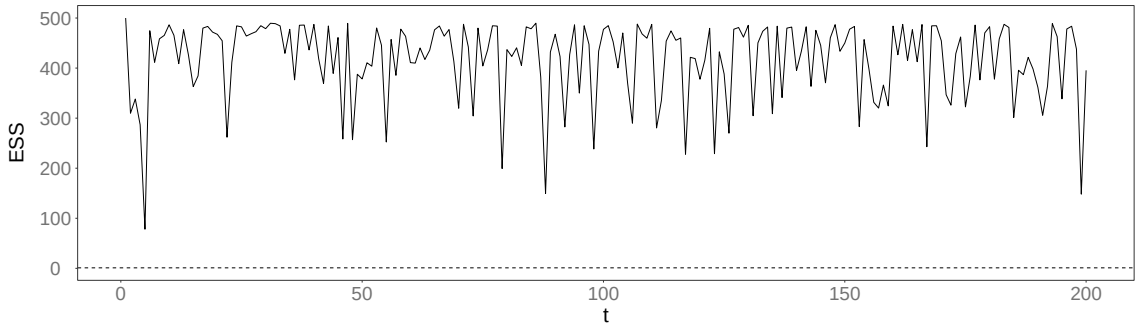


Figure 7.7: $\widehat{ESS}$ for SIR in a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with $N_{obs} = 200$ and $N_p = 500$.

density) is the original version of the SIR filter as presented by Gordon *et al.* (1993) and is usually called the *bootstrap filter* in the literature, constituting one of the most prevalent methods for state estimation in state-space models, mainly due to its simplicity and straightforward implementation.

Example. SIR state estimation for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM.

Although an analytical solution, the KF, exists for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM, to illustrate the working of the SIR filter we will implement two different variants: a SIR using the state's prior as proposal (usually called in the literature the *Bootstrap filter*, BF) and using the KF recursions as an optimal proposal (also known as a *fully adapted* SIR or SIR-FA).

Considering the model's full specification of Section 2.3.1, we have, for the BF, the following proposal:

$$\begin{aligned} \pi(\theta_t|\theta_{t-1}, y_t) &= p(\theta_t|\theta_{t-1}) \\ &= \mathcal{N}\left(\theta_t; \theta_{t-1}, \tau^2\right). \end{aligned}$$

Regarding the importance weights we will have

$$\begin{aligned} w_t &\propto w_{t-1} p\left(y_t|\theta_t\right) \\ &= w_{t-1} \mathcal{N}\left(y_t; \theta_t, \sigma^2\right). \end{aligned}$$

For the SIR-FA we can use the result in Section 6.2.1 and simplify for the $\mathcal{P}(1)$ case, that is

$$K = \tau^2 \left(\tau^2 + \sigma^2\right)^{-1}$$

and for the optimal proposal density we will have

$$\begin{aligned} \pi\left(\theta_t|\theta_{t-1}, y_t\right) &= p\left(\theta_t|\theta_{t-1}, y_t\right) \\ &= \mathcal{N}\left(\theta_t; (1-K)\,\theta_{t-1} + K y_t, (1-K)\,\tau^2\right) \\ &= \mathcal{N}\left(\theta_t; \left(1 - \frac{\tau^2}{\tau^2+\sigma^2}\right)\theta_{t-1} + \frac{\tau^2 y_t}{\tau^2+\sigma^2}, \left(1 - \frac{\tau^2}{\tau^2+\sigma^2}\right)\tau^2\right). \end{aligned}$$

As for the importance weights, and according to (6.14), we have

$$\begin{aligned} w_t &\propto w_{t-1} p\left(y_t|\theta_{t-1}\right) \\ &= w_{t-1} \mathcal{N}\left(y_t; \theta_{t-1}, \tau^2 + \sigma^2\right). \end{aligned}$$

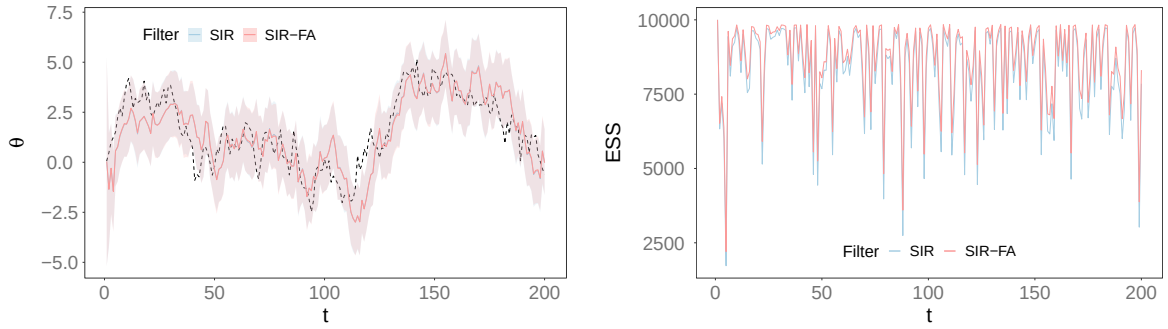
For this example we used a NDLM model with a parameter set $\boldsymbol{\Phi} = \left\{\tau^2, \nu^2\right\} = \{0.3, 4.3\}$. The resampling method used was Multinomial resampling with static checkpoint $n = 1$ (resampling at every time step). We can see the state estimation for each filter, along with the true state, for a $N_{obs} = 200$ realisation of this model in Figure 7.8a and the respective $\widehat{ESS}$ in Figure 7.8b. For this particular example, the mean $\overline{ESS}$ for SIR and SIR-FA was respectively 8340.6 and 8658.1. The MSE for the state estimation, calculated using

$$MSE_\theta = \frac{1}{N_{obs}} \sum_{t=1}^{N_{obs}} \left(\theta_t - \overline{\theta}_t\right)^2,$$

where θ_t is the "true" state from the realisation and $\overline{\theta}_t$ is the mean state value from the filter's estimation, was respectively, for SIR and SIR-FA, 1.268 and 1.266.

Example. SIR state estimation for a Poisson DLM

For this example we will compare the state estimation for a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(200, 1)\}$ Poisson DLM with the SIR filter using the prior (bootstrap filter, BS) and the CF (SIR-CF)



(a) $\theta_{0:t}$ (b) $\widehat{ESS}$

Figure 7.8: Posterior state estimation (*left*, colour line represents the posterior mean and shaded area the 90% equitailed credibility interval) and ESS (*right)* for $t = 1, \ldots, N_{obs}$ from a SIR/SIR-FA filter for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with $N_{obs} = 200$, $N_p = 10000$ and $\boldsymbol{\Phi} = \{\tau^2, \nu^2\} = \{0.3, 4.3\}$.

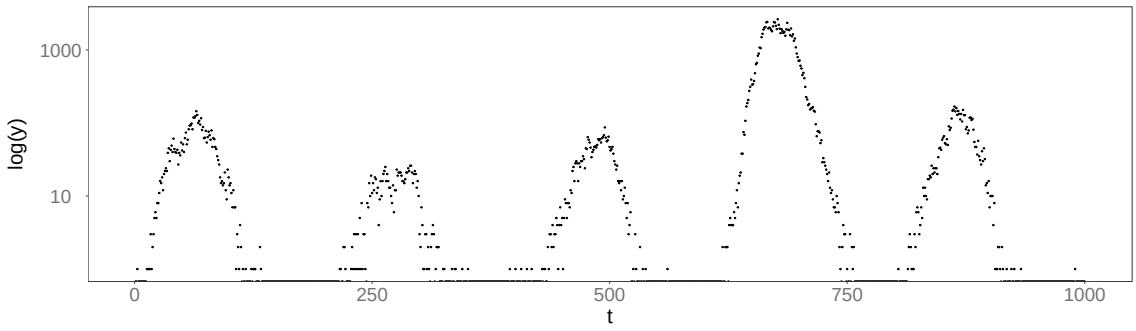


Figure 7.9: Observations for a realisation of a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(200, 1)\}$ Poisson DLM with $N_{obs} = 1000$ (*log scale*)

as proposal densities. The model can be written in the form

$$\begin{aligned} y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi} &\sim \text{Po}(\eta_t) \\ \eta_t &= \exp\left\{\mathsf{F}^T\boldsymbol{\theta}_t\right\} \\ \boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi} &\sim \mathcal{N}(\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}). \end{aligned}$$

The observation and state matrices will be respectively

$$\mathsf{F} = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^T, \qquad \mathsf{G} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sin\frac{2\pi}{p} & \cos\frac{2\pi}{p} \\ 0 & -\cos\frac{2\pi}{p} & \sin\frac{2\pi}{p} \end{bmatrix},$$

with $p = 100$. We can see in Figure 7.9 a realisation of $N_{obs} = 1000$ observations for this

Filter	MSE			time	
	θ_1	θ_2	θ_3	*total (s)*	*iteration (ms)*
BS	0.6441	1.0543	1.7485	1.093	1.09
SIR-CF	0.4580	0.5963	1.0734	3.115	3.11

Table 7.1: State posterior mean estimation MSE and computational times (total and per time step) for the BS and SIR-CF filters.

model. The SIR filter was used to perform state estimation using the prior, $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$, and the CF proposal (as described in Section 6.2.4). The estimation results for each state component are represented in Figure 7.10 on page 103. The execution times for the BS and SIR-CF were respectively 1.093 and 3.115 seconds in total (1.09 and 3.11 milliseconds per iteration). The MSE of each implementation, calculated relatively to the "true" state of model's realisation can be seen in Table 7.1. From this table we can see that, in this simulation, SIR-CF achieves a higher accuracy in estimating the states when compared to using the prior as a proposal. The calculation of the CF recursions at each step has obviously an impact in terms of computational cost, making it much higher. This is, however, a comparison made over a simple example model with simulated data and in the context on state estimation with known parameters. To better establish the behaviour of the CF proposal with more challenging data (*e.g.* possible outliers) we will use "real world" datasets in Part V on page 190 and apply the CF importance density as a proposal embedded in other filters, namely when estimating the joint posterior of states and parameters.

7.5 Particle impoverishment

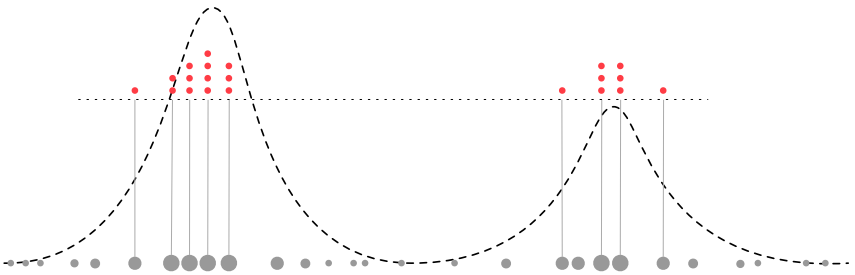


Figure 7.11: Illustration of particle impoverishment due to resampling. Grey circles represent particles at time $t-1$ (with size proportional to the weight) and dashed black line the approximated posterior density. Red circles represent the particles at time t after resampling (with uniform weights) and stacked vertically according to particle duplication. These are concentrated in regions of higher likelihood, but the diversity has diminished.

Although resampling indeed provides good results in delaying weight degeneracy, it does introduce a new problem, namely *particle impoverishment*. By selecting particles with high weights and discarding those with low weights, we are reducing the particle diversity. As we can see in Figure 7.11 a number of "surviving" particles will share their ancestry with a common pre-resampling particle. If we consider the sequence of discrete state approximation trajectories, $\left\{\boldsymbol{\theta}_{0:t}^{(i)}, w_{0:t}^{(i)}\right\}_{i=1}^{N_p}$, this implies that the entire trajectory tree is being pruned of trajectories as we go back in time. Although, by using resampling, we are keeping the particles that more closely represent the posterior, we are simultaneously using a coarser approximation by discarding particles with low weight that also help define the shape of the posterior. This effect has a clear impact on long-running SMC estimations. Of course, as in SIS, there is always the possibility of employing a "brute force" strategy and increase N_p, however this strategy is also limited by the computational cost constraints of real-time state estimation.

A result from Douc *et al.* (2005) also shows that resampling itself, being a random process, introduces additional Monte Carlo noise in the filter posterior estimates, a reason by which resampling should be applied when justified, for instance by monitoring weight degeneracy via the ESS as detailed in previous sections.

7.6 Auxiliary Particle Filter (APF)

The Auxiliary Particle Filter (APF) was first introduced by Pitt & Shephard (1999) to help mitigate the problem of particle degeneracy. One of the problems with previously mentioned methods, such as the bootstrap filter, is the fact that, when sampling the

states from the prior, $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ (6.15) we are making proposals for $\boldsymbol{\theta}_t$ without taking into account the currently available data y_t (Pitt & Shephard, 1999). This can be especially problematical in situations where outliers are present as the importance density might be a poor fit to the target density as illustrated in Figure 6.2.

The main concept behind the APF is to select particles, before propagation, according to a predictive likelihood $p\left(y_t|\boldsymbol{\theta}_{0:t-1}^{(i)}\right)$ as the criteria. The algorithm in turn is flexible in regard to this, since we can directly use it when analytically available (such as, in our context, the NDLM) or we can approximate it when not tractable (non-linear DGLMs). In the former case, the filter is usually called *fully adapted* in the literature. In cases where the predictive likelihood is not available in a closed form an approximation can be used to apply the APF algorithm. Particle filters which perform this adaptation step (reversing the order of sampling and resampling relatively to SIR) are usually said to belong to the *resample-sample* category.

If we consider the state density (3.9) and decompose it using the *predictive state density* (3.10) we have

$$\begin{aligned} p(\boldsymbol{\theta}_t|\mathcal{D}_t) &\propto p(y_t|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) \\ &\propto p(y_t|\boldsymbol{\theta}_t) \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})\, p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1})\, d\boldsymbol{\theta}_{t-1} \end{aligned} \tag{7.5}$$

From the previous results in Chapter 6 on page 66 we can use the approximation to the (empirical) *filtering density* (6.8) such that

$$p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) \approx \sum_{i=1}^{N_p} w_{t-1}^{(i)} \delta\left(\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_{t-1}^{(i)}\right).$$

By replacing with the approximation in (7.5) we have

$$p(\boldsymbol{\theta}_t|\mathcal{D}_t) \propto p(y_t|\boldsymbol{\theta}_t) \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})\, p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1})\, d\boldsymbol{\theta}_{t-1} \tag{7.6}$$

$$\approx p(y_t|\boldsymbol{\theta}_t) \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) \sum_{i=1}^{N_p} w_{t-1} \delta\left(\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_{t-1}^{(i)}\right) d\boldsymbol{\theta}_{t-1} \tag{7.7}$$

$$\propto \sum_{i=1}^{N_p} w_{t-1}^{(i)} p(y_t|\boldsymbol{\theta}_t)\, p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(i)}\right). \tag{7.8}$$

Since the empirical prediction density, $\hat{p}(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})$, has the form of a mixture of densities, Pitt & Shephard (1999) proposes sampling from a joint density $p(\boldsymbol{\theta}_t, k|\mathcal{D}_t)$ where k is

an index of said mixture. To do so, the joint density is defined as

$$\begin{aligned} p\left(\boldsymbol{\theta}_t, k|\mathcal{D}_t\right) &\propto p\left(y_t|\boldsymbol{\theta}_t\right) p\left(\boldsymbol{\theta}_t, k|\mathcal{D}_{t-1}\right) \\ &= p\left(y_t|\boldsymbol{\theta}_t\right) p\left(\boldsymbol{\theta}_t|k, \mathcal{D}_{t-1}\right) p\left(k|\mathcal{D}_{t-1}\right) \\ &= p\left(y_t|\boldsymbol{\theta}_t\right) w_{t-1}^{(k)} p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(k)}\right). \end{aligned} \tag{7.9}$$

As noted by Pitt & Shephard (1999), by sampling from $p\left(\boldsymbol{\theta}_t, k|\mathcal{D}_t\right)$ and discarding the mixture index k, we are drawing from $\hat{p}\left(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}\right)$ and then we can approximate (7.8) by

$$\hat{p}\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right) \propto \sum_{i=1}^{N_p} w_{t-1}^{(i)} p\left(y_t|\boldsymbol{\theta}_t, k^{(i)}\right) p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(i)}, k^{(i)}\right).$$

The set $\left\{\boldsymbol{\theta}_t^{(i)}, k^{(i)}\right\}_{i=1}^{N_p}$ can then be drawn from a proposal $g\left(\boldsymbol{\theta}_t, k|\mathcal{D}_t\right)$ by approximating (7.9) with

$$\begin{aligned} g\left(\boldsymbol{\theta}_t, k|\mathcal{D}_t\right) &\propto p\left(\boldsymbol{\theta}_t, k|\mathcal{D}_t\right) \\ &= p\left(y_t|\boldsymbol{\theta}_t\right) w_{t-1}^{(k)} p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(k)}\right) \\ &\approx p\left(y_t|\boldsymbol{\mu}_t^{(k)}\right) w_{t-1}^{(k)} p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(k)}\right), \end{aligned}$$

where $\boldsymbol{\mu}_t^{(k)}$ is any characterisation of $\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t^{(k)}$. According to Pitt & Shephard (1999), this is left to our design and usually is the mean, the mode or a sample. This leaves us now with the method to choose k. We factorise $g\left(\boldsymbol{\theta}_t, k|\mathcal{D}_t\right)$ as

$$g\left(\boldsymbol{\theta}_t, k|\mathcal{D}_t\right) \propto g\left(k|\mathcal{D}_t\right) g\left(\boldsymbol{\theta}_t|k, \mathcal{D}_t\right), \tag{7.10}$$

and take

$$g\left(k|\mathcal{D}_t\right) \propto w_{t-1}^{(k)} p\left(y_t|\boldsymbol{\mu}_t^{(k)}\right). \tag{7.11}$$

The indices k can then be sampled proportionally to a quantity $\lambda^{(k)}$ (called the *auxiliary weights*) such that

$$\begin{aligned} \lambda_{t-1}^{(k)} &\propto g\left(k|\mathcal{D}_t\right) \\ &= w_{t-1}^{(k)} p\left(y_t|\boldsymbol{\mu}_t^{(k)}\right). \end{aligned} \tag{7.12}$$

In conjunction with

$$g\left(\boldsymbol{\theta}_t|k^{(i)}, \mathcal{D}_t\right) = p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{\left(k^{(i)}\right)}\right), \tag{7.13}$$

this gives the necessary quantities to pre-select the particles. Pitt & Shephard (1999)

demonstrates that in order to complete the approximation of the filtering density we must calculate the adjusted weights (or *second-stage weights*) which is achieved by

$$w_t^{(i)} = w_t^{\left(k^{(i)}\right)} \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right) p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{\left(k^{(i)}\right)}\right)}{g\left(\boldsymbol{\theta}_t^{(i)}, k^{(i)}|\mathcal{D}_t\right)}$$
$$\propto \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right)}{p\left(y_t|\boldsymbol{\mu}_t^{\left(k^{(i)}\right)}\right)}.$$

Although the original method presented in Pitt & Shephard (1999) contemplated a second stage weighting and resampling we will use a version (outlined in Algorithm 7.3) which omits it, since it is not necessary and according to Carpenter *et al.* (1999) outperforms the original implementation. The general algorithm for the APF is presented in Algorithm 7.3.

Algorithm 7.3 Auxiliary Particle Filter

initialisation

for $t \leftarrow 1$ to N_{obs}

  for $i \leftarrow 1$ to N_p

    calculate $\boldsymbol{\mu}_t^{(i)}$

    calculate $\lambda_t^{(i)} \propto w_{t-1}^{(i)} p\left(y_t|\boldsymbol{\mu}_t^{(i)}\right)$

  normalise weights $\tilde{\lambda}_t^{(i)} = \frac{\lambda_t^{(i)}}{\sum_{i=1}^{N_p} \lambda_t^{(i)}}$

  resample according to $p\left(k(i) = l\right) = \tilde{\lambda}_t^{(i)}$

  for $i \leftarrow 1$ to N_p

    draw $\boldsymbol{\theta}_t^{(i)} \sim p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{\left(k^{(i)}\right)}\right)$

    calculate weights $w_t^{(i)} = \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)}\right)}{p\left(y_t|\boldsymbol{\mu}_t^{\left(k^{(i)}\right)}\right)}$

  normalise weights $\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{i=1}^{N_p} w_t^{(i)}}$

Example. Comparing the APF and SIR estimations for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM.

For this example a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM was used with a parameter set $\Phi = \left\{\tau^2, \nu^2\right\} =$

$\{0.3, 4.3\}$, corresponding to the formulation

$$\begin{aligned} y_t|\theta_t, \boldsymbol{\Phi} &\sim \mathcal{N}(\theta_t, 4.3) \\ \theta_t|\boldsymbol{\Phi} &\sim \mathcal{N}(\theta_{t-1}, 0.3). \end{aligned}$$

The initial state priors were $\theta_0 \sim \mathcal{N}(0.3, 100)$ and state estimation was performed using a bootstrap filter, as described in Section 7.4 and the APF, as described in Section 7.6. Both filters used $N_p = 500$ particles applied to $N_{obs} = 200$ observations from a realisation of the model. Multinomial resampling was used with a static checkpoint of $n = 1$, that is, performing resampling at every time step t. The results of the state estimation for both filters can be viewed in Figure 7.12a and the $\widehat{ESS}$ history can be viewed in Figure 7.12b.

For this particular example, the mean $\overline{ESS}$ for SIR and APF was respectively 417.38 and 443.71. The MSE for the state estimation was respectively, for SIR and APF, 1.30 and 1.25. The computational time was respectively, for SIR and APF, 91 and 115 milliseconds.

7.7 Forecasting

Given the state's posterior estimation up to time t, $p(\boldsymbol{\theta}_t|\mathcal{D}_t)$, SMC methods allow us to perform state and observation k-step ahead forecasting for $k \geq 1$ steps, that is $\widehat{p}(\boldsymbol{\theta}_{t+k}|\mathcal{D}_t)$. As suggested by Doucet *et al.* (2000) the forecast states can be simulated simply be propagating forward from the state model, as detailed in Algorithm 7.4.

The proposal density $\pi(\boldsymbol{\theta}_{t+n}|\boldsymbol{\theta}_{0:t+n-1}, \mathcal{D}_t)$ must be chosen, however, in a way that it doesn't depend on the value of $y_{t+1:t+k}$, a typical choice being sampling from the prior, $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$. To perform observation forecasts, that is $\hat{y}_{t:t+k}$, a similar process is applied as in the state forecast, but this time, at each forecast step, simulating observations from the observational model as described in Algorithm 7.5. In Algorithm 7.5, the observational model denoted by $f(\cdot)$ refers to any of the exponential family observational distributions considered in this thesis, *i.e.* Normal, Poisson, Binomial, etc.

Algorithm 7.2 Sequential Importance Resampling

initialisation

set state prior $\boldsymbol{\theta}_0^{(i)} \sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0)$, for $i = 1, \ldots, N_p$.

for $t \leftarrow 1$ to N_{obs}

for $i \leftarrow 1$ to N_p

Draw $\boldsymbol{\theta}_t^{(i)} \sim \pi(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t)$

Calculate the *importance weight*:

$$w_t^{(i)} = \tilde{w}_{t-1}^{(i)} \frac{p\left(y_t | \boldsymbol{\theta}_t^{(i)}\right) p\left(\boldsymbol{\theta}_t^{(i)} | \boldsymbol{\theta}_{t-1}^{(i)}\right)}{\pi\left(\boldsymbol{\theta}_t^{(i)} | \boldsymbol{\theta}_{0:t-1}^{(i)}, \mathcal{D}_t\right)}$$

Normalise weights

$$\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{i=1}^{N_p} w_t^{(i)}}$$

Calculate the *effective sample size*, $\widehat{ESS_t}$ as per (7.1).

if $\widehat{ESS_t} \leq N_{eff}$

Resample according to $p(k(i) = l) = \tilde{w}_t^{(i)}$

for $i \leftarrow 1$ to N_p

set weights $\tilde{w}_t^{(i)} = \frac{1}{N_p}$

assign states $\boldsymbol{\theta}_t^{(i)} = \boldsymbol{\theta}_t^{\left(k^{(i)}\right)}$

calculate state approximation

$$p(\boldsymbol{\theta}_t | \mathcal{D}_t) \approx \sum_{i=1}^{N_p} \tilde{w}_t^{(i)} \delta\left(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(i)}\right)$$

Algorithm 7.4 State forecasting

for $m \leftarrow 1$ to k

for $n \leftarrow 1$ to N_p

sample $\boldsymbol{\theta}_{t+n}^{(i)} \sim \pi\left(\boldsymbol{\theta}_{t+n} | \boldsymbol{\theta}_{0:t+n-1}^{(i)}, \mathcal{D}_t\right)$

set

$$p(\boldsymbol{\theta}_{t+n} | \mathcal{D}_t) \approx \frac{1}{N_p} \sum_{i=1}^{N_p} \delta\left(\boldsymbol{\theta}_{t+n} - \boldsymbol{\theta}_{t+n}^{(i)}\right)$$

Algorithm 7.5 State forecasting

for $m \leftarrow 1$ to k

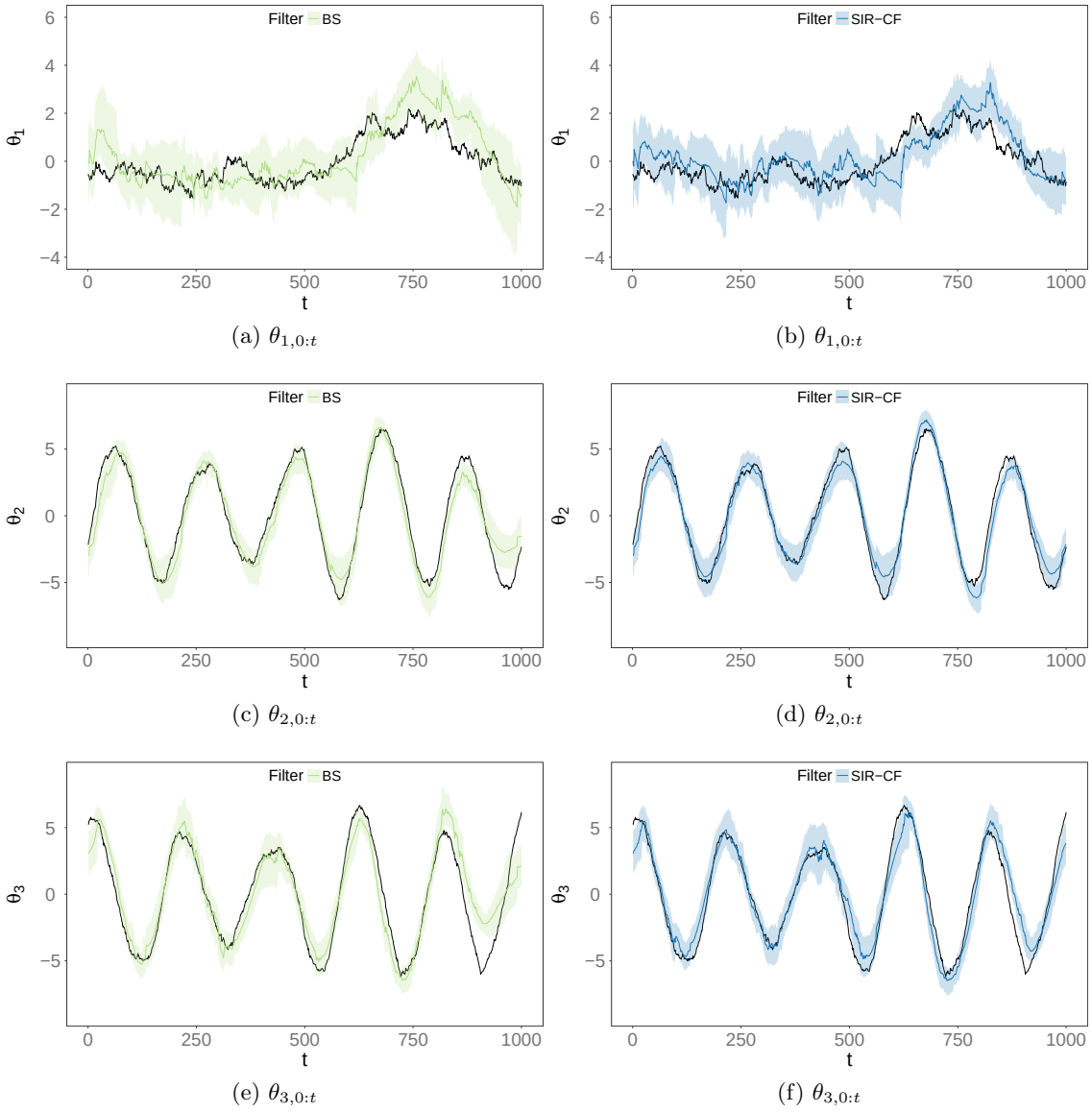
for $n \leftarrow 1$ to N_p

sample $\boldsymbol{\theta}_{t+n}^{(i)} \sim \pi\left(\boldsymbol{\theta}_{t+n} | \boldsymbol{\theta}_{0:t+n-1}^{(i)}, \mathcal{D}_t\right)$

sample $y_{t+n}^{(i)} | \boldsymbol{\theta}_{t+n} \sim f\left(y_t | \boldsymbol{\theta}_{t+n}\right)$

set

$$\begin{aligned} p\left(\boldsymbol{\theta}_{t+n} | \mathcal{D}_t\right) &\approx \frac{1}{N_p} \sum_{i=1}^{N_p} \delta\left(\boldsymbol{\theta}_{t+n} - \boldsymbol{\theta}_{t+n}^{(i)}\right) \\ p\left(y_{t+n} | \boldsymbol{\theta}_{t+n}\right) &\approx \frac{1}{N_p} \sum_{i=1}^{N_p} \delta\left(y_{t+n} - y_{t+n}^{(i)}\right) \end{aligned}$$

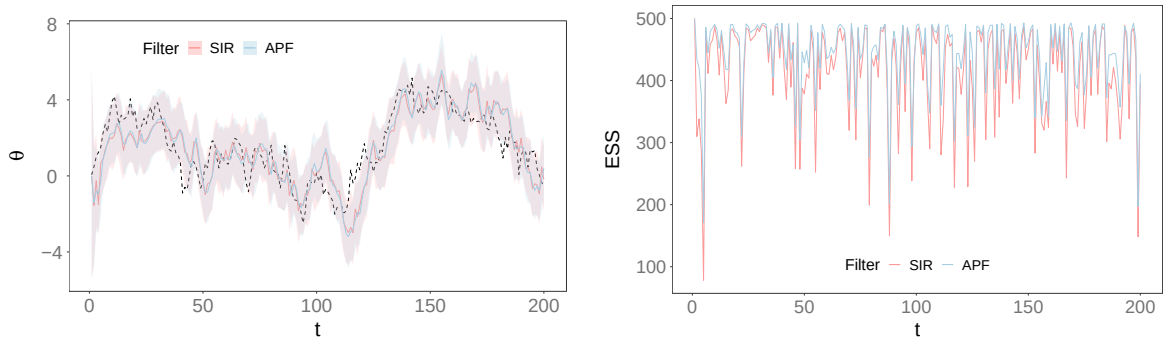


(a) $\theta_{1,0:t}$ (b) $\theta_{1,0:t}$

(c) $\theta_{2,0:t}$ (d) $\theta_{2,0:t}$

(e) $\theta_{3,0:t}$ (f) $\theta_{3,0:t}$

Figure 7.10: Estimation of the three state components $(\theta_1, \theta_2, \theta_3)$ for a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(200, 1)\}$ Poisson DLM, using the BS filter (*left column*) and SIR-CF (*right column*). Black line represent the "true" state, colour line is the state posterior mean and shaded area the 90% equitailed credibility intervals.



(a) $\theta_{0:t}$ (b) $\widehat{ESS}$

Figure 7.12: State estimation (*left*) and ESS (*right*) for $N_{obs} = 200$ observations for a SIR filter and APF for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with $N_{obs} = 200$ and $N_p = 500$.

. On the left, solid colour lines represent the state posterior mean and shaded areas the 90% equitailed credibility intervals. Dashed black line represents the true state posterior mean.

Part III

Online State and Parameter Estimation

Chapter 8

State augmentation approaches

So far, we have been focusing on methods to estimate the state in DGLMs given a static and known parameter set $\boldsymbol{\Phi}$ as described in Section 2.1 on page 4, that is

$$p\left(\boldsymbol{\theta}_{0:t}|\mathcal{D}_t, \boldsymbol{\Phi}\right).$$

However, in real world scenarios, both the states and parameters will be unknown and we will be interested in the estimation of the joint distribution

$$p\left(\boldsymbol{\theta}_{0:t}, \boldsymbol{\Phi}|\mathcal{D}_t\right).$$

Initial work on state-space state estimation methods included the concept of propagating samples using "shocks" and applying "roughening penalties" (Liu & West, 2001) which lead to standard (Gordon *et al.*, 1993) particle filtering. Historically, the concepts introduced in Gordon *et al.* (1993) for state estimation (described in Chapter 7.4), provided a natural basis which could be extended to the problem of joint state and parameter estimation in state-space models. Early techniques tried to solve this problem by incorporating the parameters in the state-space and applying standard SMC. The state-space could then be explored by propagating the states while at the same time adding artificial noise to the parameters. However, since in many scenarios we are dealing with static parameters, a problem with this approach is that static parameters, by definition, will not change their value from $t = 1$. If we incorporate the parameters as part of the state-space, artificial noise will then inevitably aggravate any degeneracy problems and also lead to an artificial increase variance of both the state and parameter estimates. Degeneracy problems are to be expected due to inevitable collapse of the parameter component of the augmented state-space.

Several methods were proposed to alleviate this problem, such as self-organising state-

space models, as initially introduced by Kitagawa (1998). Additional proposed methods included the usage of kernel density estimation methods, although these methods usually incur in an artificial increase in the parameters' posterior variances. An extension of such methods, the *Liu and West* particle filter, first described in Liu & West (2001), tries to solve this particular problem and is described in this chapter.

The solution proposed by the Liu and West filter (Liu & West, 2001) is to use a kernel-smoothing approximation with a correction factor to account for over-dispersion.

If the parameter set $\boldsymbol{\Phi}$ is considered as being time-evolving (that is as $\boldsymbol{\Phi}_t$)[1] and we also consider $\boldsymbol{\Phi}_t$ as evolving according to zero-mean, normally distributed increments, we can write, according to Liu & West (2001)

$$\begin{aligned} \boldsymbol{\Phi}_t &= \boldsymbol{\Phi}_{t-1} + \boldsymbol{\zeta}_t \\ \boldsymbol{\zeta}_t &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_t). \end{aligned} \tag{8.1}$$

Assuming $\boldsymbol{\Phi}_{t-1} \perp\!\!\!\perp \boldsymbol{\zeta}_t | \mathcal{D}_t$, this would then enable the state-space to be extended with the parameter set $\boldsymbol{\Phi}_t$ and performing online state estimation as described in the previous chapters.

By adding artificial noise we are delaying the particle degeneracy that would otherwise occur, but in a quicker way, by simply incorporating a static $\boldsymbol{\Phi}$ in the state-space. According to Liu & West (2001), this artificial noise approach has the problem of incurring information loss and originating a far greater variance to the parameter estimation than the "true" posteriors.

Liu & West (2001) proposes to reinterpret the parameter noise term in the context of *Kernel Smoothing* (West, 1993). If at a time t we have a discrete approximation of the parameter posterior as $\boldsymbol{\Phi}_t$ as a set of parameter particles with their respective weights, $\{\boldsymbol{\Phi}_t^{(i)}, w_t^{(i)}\}_{i=1}^{N_p}$, similarly to the discrete state approximation in Section 6 on page 66, the mean and variance can be defined as

$$\overline{\boldsymbol{\Phi}}_t = \sum_{i=1}^{N_p} w_t^{(i)} \boldsymbol{\Phi}_i \tag{8.2}$$

$$\mathsf{V}_t = \sum_{i=1}^{N_p} w_t^{(i)} \left(\boldsymbol{\Phi}_i - \overline{\boldsymbol{\Phi}}_t\right) \left(\boldsymbol{\Phi}_i - \overline{\boldsymbol{\Phi}}_t\right)^T. \tag{8.3}$$

By taking $\left\{\mathbf{m}_t^{(i)}\right\}_{i=1}^{N_p}$ as the kernel locations and $h > 0$ as a smoothing parameter, then,

[1] Note that the notation $\boldsymbol{\Phi}_t$ means that we are considering a time-variant parameter. We will, in later chapters, introduce a similar notation but signifying the posterior estimate of $\boldsymbol{\Phi}$ at time t.

according to West (1993), the smoothed kernel density will be

$$p\left(\boldsymbol{\Phi}|\mathcal{D}_t\right) \approx \sum_{i=1}^{N_p} w_t^{(i)} \mathcal{N}\left(\boldsymbol{\Phi}|\mathbf{m}_t^{(i)}, h^2\mathsf{V}_t\right). \tag{8.4}$$

According to Liu & West (2001), by using the standard kernel methods in (8.4), where the kernel locations are assumed as the sample values $\mathbf{m}_t^{(i)} = \boldsymbol{\Phi}_t^{(i)}$, we would have a mixture variance of $\left(1+h^2\right)\mathsf{V}_t$, always larger than V_t. Since we know that for the mixture of normals in (8.4):

$$\begin{aligned}
\mathrm{E}\left[\boldsymbol{\Phi}|\mathcal{D}_t\right] &= \sum_{i=1}^{N_p} w_t^{(i)} \mathbf{m}_t^{(i)} \\
\mathrm{Var}\left[\boldsymbol{\Phi}|\mathcal{D}_t\right] &= \sum w_t^{(i)} \left\{ \left(\mathbf{m}_t^{(i)} - \overline{\mathbf{m}}_t\right)\left(\mathbf{m}_t^{(i)} - \overline{\mathbf{m}}_t\right)^T + h^2\mathsf{V}_t \right\}
\end{aligned}$$

by taking $\mathbf{m}_t^{(i)} = \boldsymbol{\Phi}_t^{(i)}$ we have

$$\begin{aligned}
\mathrm{E}\left[\boldsymbol{\Phi}|\mathcal{D}_t\right] &= \sum_{i=1}^{N_p} w_t^{(i)} \boldsymbol{\Phi}_t^{(i)} = \overline{\boldsymbol{\Phi}}_t \\
\mathrm{Var}\left[\boldsymbol{\Phi}|\mathcal{D}_t\right] &= \sum_{i=1}^{N_p} w_t^{(i)} \left\{ \underbrace{\left(\boldsymbol{\Phi}_t^{(i)} - \overline{\boldsymbol{\Phi}}_t\right)\left(\boldsymbol{\Phi}_t^{(i)} - \overline{\boldsymbol{\Phi}}_t\right)^T}_{\mathsf{V}_t} + h^2\mathsf{V}_t \right\} \\
&= \sum_{i=1}^{Np} w_t^{(i)} \left\{\mathsf{V}_t + h^2\mathsf{V}_t\right\} = \sum_{i=1}^{Np} w_t^{(i)} \left\{\left(1+h^2\right)\mathsf{V}_t\right\} = \left(1+h^2\right)\mathsf{V}_t,
\end{aligned}$$

which, as mentioned, will lead to an over-dispersion of the kernel density in comparison to the posterior.

Liu & West (2001) introduces the concept of *kernel location shrinkage* by using the new locations

$$\mathbf{m}_t^{(i)} = a\boldsymbol{\Phi}_t^{(i)} + (1-a)\,\overline{\boldsymbol{\Phi}}_t,$$

where $a = \sqrt{1-h^2}$ is the shrinkage parameter. The resulting posterior mean, using

shrinkage, will then be

$$\begin{aligned}
\mathrm{E}\left[\boldsymbol{\Phi}|\mathcal{D}_t\right] &= \sum_{i=1}^{N_p} w_t^{(i)} \left(a\boldsymbol{\Phi}_t^{(i)} + (1-a)\,\overline{\boldsymbol{\Phi}}_t\right) \\
&= \sum_{i=1}^{N_p} w_t^{(i)} \left\{\sqrt{1-h^2}\boldsymbol{\Phi}_t^{(i)} + \left(1-\sqrt{1-h^2}\right)\overline{\boldsymbol{\Phi}}_t\right\} \\
&= \sqrt{1-h^2}\underbrace{\sum_{i=1}^{N_p} w_t^{(i)}\boldsymbol{\Phi}_t^{(i)}}_{\overline{\Phi}_t} + \left(1-\sqrt{1-h^2}\right)\overline{\boldsymbol{\Phi}}_t\underbrace{\sum_{i=1}^{N_p} w_t^{(i)}}_{1} \\
&= \sqrt{1-h^2}\overline{\boldsymbol{\Phi}}_t + \left(1-\sqrt{1-h^2}\right)\overline{\boldsymbol{\Phi}}_t = \overline{\boldsymbol{\Phi}}_t, \qquad (8.5)
\end{aligned}$$

whereas the posterior variance will be

$$\begin{aligned}
\mathrm{Var}\left[\boldsymbol{\Phi}|\mathcal{D}_t\right] &= \sum_{i=1}^{N_p} w_t^{(i)} \left\{\left(\left\{a\boldsymbol{\Phi}_t^{(i)} + (1-a)\,\overline{\boldsymbol{\Phi}}_t\right\} - \overline{\boldsymbol{\Phi}}_t\right)\left(\left\{a\boldsymbol{\Phi}_t^{(i)} + (1-a)\,\overline{\boldsymbol{\Phi}}_t\right\} - \overline{\boldsymbol{\Phi}}_t\right)^T + h^2\mathsf{V}_t\right\} \\
&= \sum_{i=1}^{N_p} w_t^{(i)} \left\{\left(a\boldsymbol{\Phi}_t^{(i)} + a\overline{\boldsymbol{\Phi}}_t\right)\left(a\boldsymbol{\Phi}_t^{(i)} + a\overline{\boldsymbol{\Phi}}_t\right)^T + h^2\mathsf{V}_t\right\} \\
&= \sum_{i=1}^{N_p} w_t^{(i)} \left\{a^2\underbrace{\left(\boldsymbol{\Phi}_t^{(i)} + \overline{\boldsymbol{\Phi}}_t\right)\left(\boldsymbol{\Phi}_t^{(i)} + \overline{\boldsymbol{\Phi}}_t\right)^T}_{V_t} + h^2\mathsf{V}_t\right\} \\
&= \sum_{i=1}^{Np} w_t^{(i)} \left(\underbrace{a^2}_{1-h^2} + h^2\right)\mathsf{V}_t = \mathsf{V}_t. \qquad (8.6)
\end{aligned}$$

By using shrinkage we obtain the correct mean and variance for the $\boldsymbol{\Phi}_t$ posterior, as shown in (8.5) and (8.6).

Assuming the $\boldsymbol{\Phi}_{t+1}$ update by $\boldsymbol{\zeta}_{t+1}$, as described in (8.1), we know that the prior $p\left(\boldsymbol{\Phi}_{t+1}|\mathcal{D}_t\right)$ will have mean $\boldsymbol{\Phi}_t$ and variance $\mathsf{V}_t + \Omega_{t+1}$. In Liu & West (2001) the discrete approximation of $p\left(\boldsymbol{\Phi}_{t+1}|\mathcal{D}_t\right)$ is defined as

$$p\left(\boldsymbol{\Phi}_{t+1}|\mathcal{D}_t\right) \approx \sum_{i=1}^{N_p} w_t^{(i)} \mathcal{N}\left(\boldsymbol{\Phi}_{t+1}|\boldsymbol{\Phi}_t^{(i)}, \mathsf{V}_t + \Omega_{t+1}\right),$$

which is also in the kernel form. Since we know[2] the variance of the prior is

$$\begin{aligned}\text{Var}\left[\boldsymbol{\Phi}_{t+1}|\mathcal{D}_t\right] &= \text{Var}\left[\boldsymbol{\Phi}_t|\mathcal{D}_t\right] + \text{Var}\left[\boldsymbol{\zeta}_{t+1}|\mathcal{D}_t\right] + 2\text{Cov}\left[\boldsymbol{\Phi}_t, \boldsymbol{\zeta}_{t+1}|\mathcal{D}_t\right] \\ &= \text{Var}\left[\boldsymbol{\Phi}_t|\mathcal{D}_t\right] + \Omega_{t+1} + 2\text{Cov}\left[\boldsymbol{\Phi}_t, \boldsymbol{\zeta}_{t+1}|\mathcal{D}_t\right],\end{aligned}$$

we want it to be equal to the prior variance as V_t, that is

$$\text{Var}\left[\boldsymbol{\Phi}_{t+1}|\mathcal{D}_t\right] = \text{Var}\left[\boldsymbol{\Phi}_t|\mathcal{D}_t\right] = \mathsf{V}_t.$$

By setting

$$\text{Cov}\left[\boldsymbol{\Phi}_t, \boldsymbol{\zeta}_{t+1}|\mathcal{D}_t\right] = -\frac{\Omega_{t+1}}{2},$$

we then have

$$\begin{aligned}\text{Var}\left[\boldsymbol{\Phi}_{t+1}|\mathcal{D}_t\right] &= \text{Var}\left[\boldsymbol{\Phi}_t|\mathcal{D}_t\right] + \Omega_{t+1} + 2\underbrace{\text{Cov}\left[\boldsymbol{\Phi}_t, \boldsymbol{\zeta}_{t+1}|\mathcal{D}_t\right]}_{-\frac{\Omega_{t+1}}{2}} \\ &= \text{Var}\left[\boldsymbol{\Phi}_t|\mathcal{D}_t\right] + \Omega_{t+1} - 2\frac{\Omega_{t+1}}{2} \\ &= \text{Var}\left[\boldsymbol{\Phi}_t|\mathcal{D}_t\right].\end{aligned}$$

Assuming joint normality of $\left(\boldsymbol{\Phi}_t, \boldsymbol{\zeta}_{t+1}|\mathcal{D}_t\right)$,we can write $p\left(\boldsymbol{\Phi}_{t+1}|\boldsymbol{\Phi}_t\right)$ as

$$\begin{aligned}p\left(\boldsymbol{\Phi}_{t+1}|\boldsymbol{\Phi}_t\right) &= \mathcal{N}\left(\boldsymbol{\Phi}_{t+1}|\mathsf{A}_{t+1}\boldsymbol{\Phi}_t + \left(\mathbf{I} - \mathsf{A}_{t+1}\right)\overline{\boldsymbol{\Phi}}_t, \left(\mathbf{I} - \mathsf{A}_{t+1}^2\right)\mathsf{V}_t\right) \\ \mathsf{A}_{t+1} &= \mathbf{I} - \frac{\Omega_{t+1}\mathsf{V}_t^{-1}}{2}.\end{aligned}$$

In order to define the artificial noise covariance Ω_{t+1}, Liu & West (2001) defines it in terms of a discount factor, such that

$$\Omega_{t+1} = \mathsf{V}_t\left(\frac{1}{\delta} - 1\right)$$

with $0 < \delta \leq 1$, and usually chosen to be close to $\delta = 0.99$. To write the above $\boldsymbol{\Phi}_{t+1}$ evolution in the kernel form we take

$$\begin{aligned}\mathsf{A}_{t+1} &= a\mathbf{I} \\ a &= \frac{(3\delta - 1)}{2\delta}.\end{aligned}$$

We can then rewrite the evolution as (taking the relation $a = \sqrt{1 - h^2}$ and express in terms

[2] cf. Section F.1.1.

$h^2 = 1 - a^2$):

$$\begin{aligned} p\left(\boldsymbol{\Phi}_{t+1}|\boldsymbol{\Phi}_t\right) &= \mathcal{N}\left(\boldsymbol{\Phi}_{t+1}|\mathsf{A}_{t+1}\boldsymbol{\Phi}_t + \left(\mathbf{I} - \mathsf{A}_{t+1}\right)\overline{\boldsymbol{\Phi}}_t, \left(\mathbf{I} - \mathsf{A}_{t+1}^2\right)\mathsf{V}_t\right) \\ &= \mathcal{N}\left(\boldsymbol{\Phi}_{t+1}|a\mathbf{I}\boldsymbol{\Phi}_t + \left(\mathbf{I} - a\mathbf{I}\right)\overline{\boldsymbol{\Phi}}_t, \left(\mathbf{I} - a^2\mathbf{I}^2\right)\mathsf{V}_t\right) \\ &= \mathcal{N}\left(\boldsymbol{\Phi}_{t+1}|\mathbf{I}\left(a\boldsymbol{\Phi}_t + \overline{\boldsymbol{\Phi}}_t\left(1-a\right)\right), \left(1-a^2\right)\mathsf{V}_t\right) \\ &= \mathcal{N}\left(\boldsymbol{\Phi}_{t+1}|a\boldsymbol{\Phi}_t + \left(1-a\right)\overline{\boldsymbol{\Phi}}_t, \underbrace{\left(1-a^2\right)}_{h^2}\mathsf{V}_t\right) \\ &= \mathcal{N}\left(\boldsymbol{\Phi}_{t+1}|a\boldsymbol{\Phi}_t + \left(1-a\right)\overline{\boldsymbol{\Phi}}_t, h^2\mathsf{V}_t\right). \end{aligned}$$

Liu & West (2001) proposes to use this method of estimating the marginal $\boldsymbol{\Phi}_t$ density within the APF framework as described in Section 7.6.

(For N_p particles) we start by establishing the priors of $\{\boldsymbol{\theta}_t, \boldsymbol{\Phi}\}$ using $\left\{\boldsymbol{\mu}_{t+1}^{(i)}, \mathbf{m}_t^{(i)}\right\}_{i=1,\ldots,N_p}$ where, as discussed in Section 7.6, $\boldsymbol{\mu}_{t+1}^{(i)}$ will be some characterisation of $\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t^{(i)}$, such as

$$\boldsymbol{\mu}_{t+1}^{(i)} = \mathrm{E}\left[\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t^{(i)}, \Phi^{(i)}\right]$$

and $\left\{\mathbf{m}_t^{(i)} = a\boldsymbol{\Phi}_t^{(i)} + \left(1-a\right)\overline{\boldsymbol{\Phi}}_t\right\}_{i=1}^{N_p}$ are the kernel locations. As in the APF (7.12) we sample an auxiliary index k with probability

$$\lambda_{t+1}^{(i)} \propto w_t^{(i)} p\left(y_{t+1}|\boldsymbol{\mu}_{t+1}^{(i)}, \mathbf{m}_t^{(i)}\right).$$

In the LW filter we additionally sample the new proposed parameter set $\boldsymbol{\Phi}_{t+1}^{(i)}$ from the the above:

$$\boldsymbol{\Phi}_{t+1}^{(i)} \sim \mathcal{N}\left(\boldsymbol{\Phi}_{t+1}|\mathbf{m}_t^{(i)}, h^2\mathsf{V}_t\right).$$

According to Liu & West (2001); Prado & West (2010) the parameters can be transformed whenever appropriate. As an example, when dealing with parameters $\Phi \in (0,1]$, (such as in the NBDLM) it might be appropriate to work in the transformed

$$\Phi^\star = \log\left(\frac{\Phi}{1-\Phi}\right),$$

and when working with parameters $\Phi > 0$ (such as the state variances in the DGLM) we could work in the log-space such that

$$\Phi^\star = \log\left(\Phi\right).$$

The new states can be sampled from the system equation (conditioned on the new para-

meters)

$$\boldsymbol{\theta}_{t+1}^{(i)} \sim \mathcal{N}\left(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_{t+1}^{(i)}, \boldsymbol{\Phi}_{t+1}^{(i)}\right),$$

and finally we calculate the second-stage weights

$$w_{t+1}^{(i)} \propto \frac{p\left(y_{t+1}|\boldsymbol{\theta}_{t+1}^{(i)}, \boldsymbol{\Phi}_{t+1}^{(i)}\right)}{p\left(y_{t+1}|\boldsymbol{\mu}_{t+1}^{(i)}, \mathbf{m}_{t}^{(i)}\right)}.$$

This method is specifically defined to estimate fixed parameters, however the LW filter is an extremely versatile instrument for state and parameter estimation since it can applied with minimal assumptions about the structural relationship between parameters. It can also be considered as a benchmark for state and parameter estimation by SMC methods (Lopes & Carvalho, 2011). It should be noted however that incorporating the parameters in the state-space can worsen any degeneracy problems. The full general algorithm for the LW filter is presented in Algorithm 8.1.

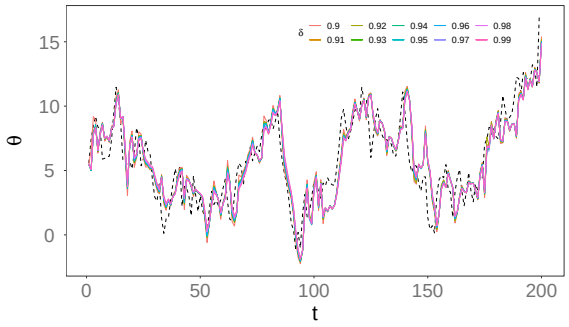
Example. Effect of smoothing parameter δ.

We consider here a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with parameters $\tau^2 = 1.5$ and $\nu^2 = 3.5$. The true states for a realisation of this model are presented in Figure 8.1a on the facing page. The initial state priors are taken as $\theta_0 \sim \mathcal{N}(5.5, 10)$ and the prior was used as the importance density. The parameters priors are taken as $\tau_0^2 \sim \mathcal{IG}(1,1)$ and $\nu_0^2 \sim \mathcal{IG}(1,1)$. The estimation was performed using $N_p = 10^5$ particles and using different values of $\delta = \{0.90, 0.91, 0.92, \ldots, 0.99\}$. Resampling was performed using Multinomial resampling with a static checkpoint of $n = 1$.

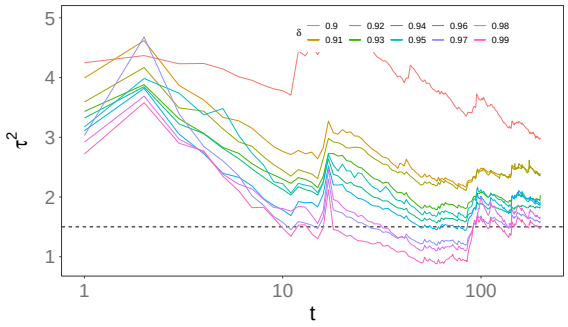
The state estimation is presented in Figure 8.1a on the next page and the parameter posterior estimation mean for τ^2 and ν^2 can be viewed in figure 8.1b and 8.1c on the facing page respectively. We can see from Figure 8.1e on the next page that, in this case, varying δ did not impact in any obvious way the posterior estimation of $p\left(\nu^2|\mathcal{D}_t\right)$. However, there is a clear pattern in the $p\left(\tau^2|\mathcal{D}_t\right)$ estimation (Figure 8.1d on the facing page), with values of δ closer to 1 leading to a posterior estimation closer to the "true" value. This is consistent with the general advice in the literature.

Example. LW on $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM using a Fully Adapted (FA) proposal

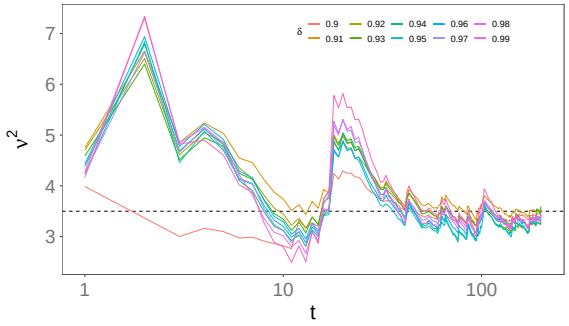
Here we consider here a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with parameters $\tau^2 = 0.75$ and $\nu^2 = 1.25$. The initial state priors are taken as $\theta_0 \sim \mathcal{N}(0.1, 10)$ and, since we are dealing with a linear model, the fully adapted version of LW was used. The parameters priors are taken as $\tau_0^2 \sim \mathcal{IG}(1,1)$ and $\nu_0^2 \sim \mathcal{IG}(1,1)$. The estimation was performed using $N_p = 10^5$, with a smoothing parameter of $\delta = 0.99$ and $N_{obs} = 200$ data points. The resampling method used was stratified with a static checkpoint of $n = 1$.



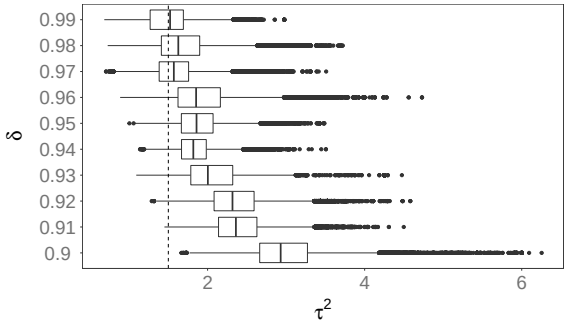
(a) LW state posterior mean estimation for different values of δ. Dashed line represents true value.



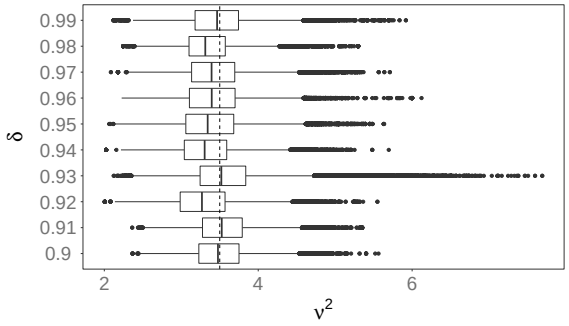
(b) LW τ^2 posterior mean estimation for different values of δ. Dashed line represents true value



(c) LW ν^2 posterior mean estimation for different values of δ. Dashed line represents true value



(d) $p\left(\tau^2|\mathcal{D}_{N_{obs}}\right)$ for different values of δ. Dashed line represents the parameter true value



(e) $p\left(\nu^2|\mathcal{D}_{N_{obs}}\right)$ for different values of δ. Dashed line represents the parameter true value

Figure 8.1: State and parameter posterior mean estimation for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM using different δ values.

Algorithm 8.1 Liu and West algorithm

initialisation

$a = \frac{(3\delta-1)}{2\delta}$

$\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0)$

$\boldsymbol{\Phi}_0 \sim p(\boldsymbol{\Phi})$

for $t \leftarrow 1$ to N_{obs}

for $i \leftarrow 1$ to N_p

calculate $\mathbf{m}_{t-1}^{(i)} = a\boldsymbol{\Phi}_{t-1}^{(i)} + (1-a)\overline{\boldsymbol{\Phi}}_{t-1}$

calculate $\boldsymbol{\mu}_t^{(i)}$

calculate $\lambda_t^{(i)} \propto w_{t-1}^{(i)} p\left(y_t | \boldsymbol{\mu}_t^{(i)}, \mathbf{m}_{t-1}^{(i)}\right)$

normalise weights $\widetilde{\lambda}_t^{(i)} = \frac{\lambda_t^{(i)}}{\sum_{i=1}^{N_p} \lambda_t^{(i)}}$

resample according to $p\left(k^{(i)} = k\right) = \widetilde{\lambda}_t^{(k)}$

calculate $\mathsf{V}_{t-1} = \sum \left(\boldsymbol{\Phi}_{t-1}^{(i)} - \overline{\boldsymbol{\Phi}}_{t-1}\right)\left(\boldsymbol{\Phi}_{t-1}^{(i)} - \overline{\boldsymbol{\Phi}}_{t-1}\right)^T w_{t-1}^{(i)}$

for $i \leftarrow 1$ to N_p

update parameters $\boldsymbol{\Phi}_t^{(i)} \sim \mathcal{N}\left(\mathbf{m}_{t-1}^{\left(k^{(i)}\right)}, \left(1-a^2\right)\mathsf{V}_{t-1}\right)$

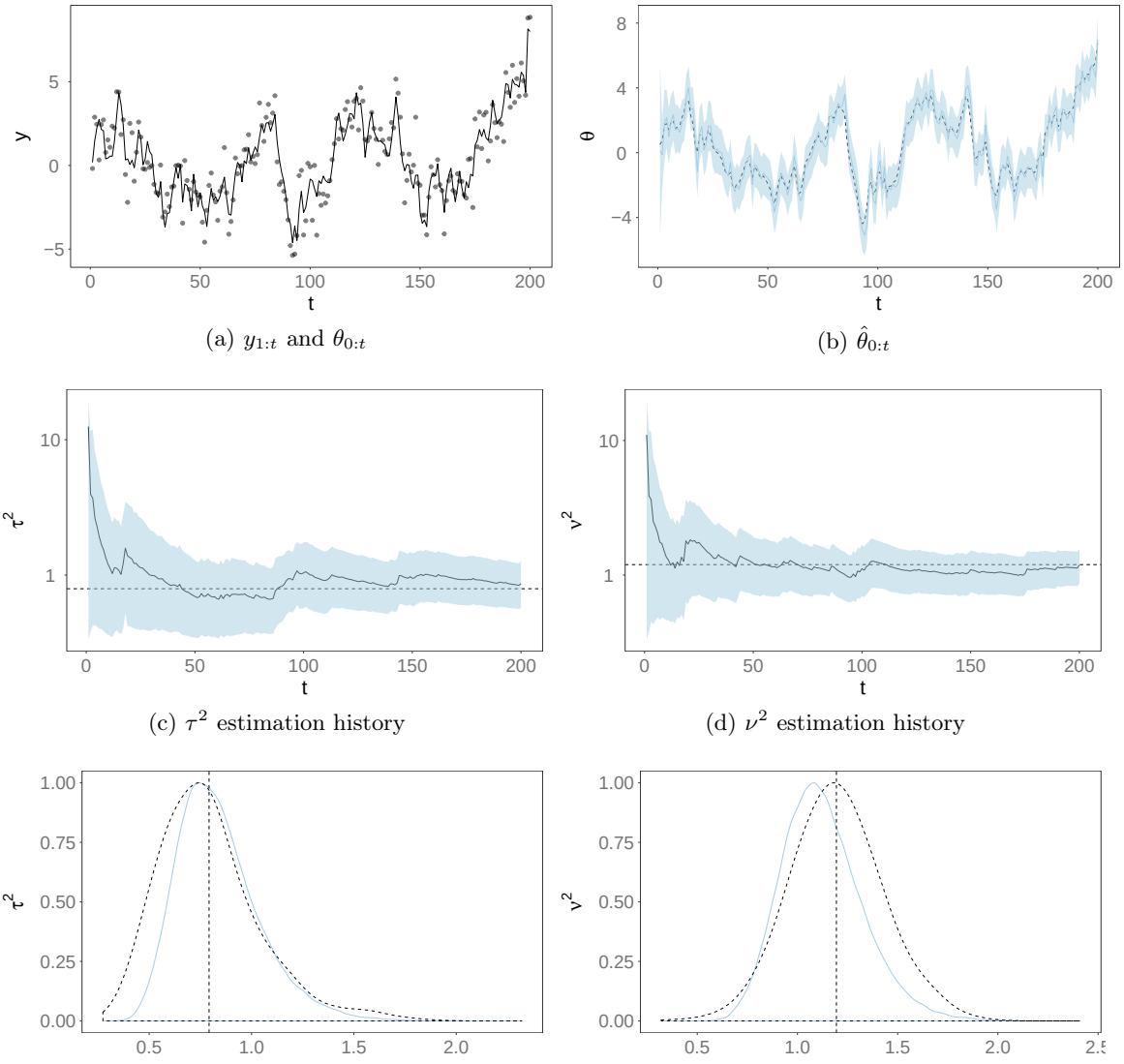
draw $\boldsymbol{\theta}_t^{(i)} \sim p\left(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}^{\left(k^{(i)}\right)}, \boldsymbol{\Phi}_t^{(i)}\right)$

calculate weights $w_t^{(i)} = \frac{p\left(y_t | \boldsymbol{\theta}_t^{(i)}, \boldsymbol{\Phi}^{(i)}\right)}{p\left(y_t | \boldsymbol{\mu}_t^{\left(k^{(i)}\right)}, \mathbf{m}_{t-1}^{\left(k^{(i)}\right)}\right)}$

normalise weights $\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^{N_p} w_t^{(i)}}$

In Figure 8.2a on the next page we can see the data and states for a realisation of this model and in Figure 8.2b on the facing page the filter's state estimation (along with 90% equitailed credibility intervals) compared to the true posterior mean[3]. In Figures 8.2c on the facing page and 8.2d on the next page we can see evolution, for $t = 0, \ldots, N_{obs}$ of $p\left(\tau^2 | \mathcal{D}_t\right)$ and $p\left(\nu^2 | \mathcal{D}_t\right)$ respectively. In Figures 8.2e and 8.2f we can see the parameters' marginals at $t = 200$ respectively for τ^2 and ν^2. The LW's posterior mean value for $p\left(\tau^2 | \mathcal{D}_T\right)$ was 0.8402 and for $p\left(\nu^2 | \mathcal{D}_T\right)$ was 1.1253 and the "true" values were, respectively, 0.7932 and 1.2011. The MSE between the "true" states and the LW estimated state posterior mean was 0.1972.

[3]State and parameter posteriors estimated offline using a long run of Particle Markov Chain Monte Carlo (PMMC). This method will be fully discussed in Chapter 13 on page 152.



(a) $y_{1:t}$ and $\theta_{0:t}$

(b) $\hat{\theta}_{0:t}$

(c) τ^2 estimation history

(d) ν^2 estimation history

(e) $p\left(\tau^2|\mathcal{D}_t\right)$ marginal at $t = 200$

(f) $p\left(\nu^2|\mathcal{D}_t\right)$ marginal at $t = 200$

Figure 8.2: State and parameter estimation using LW for a realisation of a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with $\boldsymbol{\Phi} = \left\{\tau^2, \nu^2\right\} = \{0.75, 1.25\}$. Realisation data (*dots*) and states (*line*) in 8.2a. State posterior mean as coloured line, shaded area as 90% equitailed credibility intervaland ground truth as dashed line in 8.2b. Parameter posterior estimation history with coloured line as posterior mean, shaded area as 90% equitailed credibility intervaland horizontal dashed line as truth in 8.2c and 8.2d. τ^2 and ν^2 posteriors at $t = N_{obs}$ using LW in blue and truth as dashed line in 8.2e and 8.2f.

Chapter 9

Storvik filter

Another important method for sequential joint estimation of states and static parameters is the *Storvik filter*, first presented in Storvik (2002) and built on work in Fearnhead (2002). The Storvik filter uses the sample-resample framework, *i.e.* the SIR framework of Section 7.4 on page 91. However, in contrast with the LW filter, the state-space is not augmented with the parameters, but instead the parameters will be marginalised and estimated sequentially using a finite set of values, called *sufficient statistics*. The concept of sufficient statistics, in this context, is defined as presented in Fearnhead (2002) and not necessarily in the generally used definition of sufficient statistics. In Fearnhead (2002) the use of sufficient statistics arises from the problem of applying MCMC moves to SMC. If we consider a joint distribution of states and a parameter $\boldsymbol{\Phi}$ as

$$p\left(\boldsymbol{\theta}_t, \boldsymbol{\Phi} | \mathcal{D}_t\right),$$

according to Fearnhead (2002), a stationary distribution can be designed (up to a constant) for the joint distribution of the trajectories, that is

$$p\left(\boldsymbol{\theta}_{0:t}, \boldsymbol{\Phi} | \mathcal{D}_t\right) = \pi\left(\boldsymbol{\theta}_0, \boldsymbol{\Phi}\right) \prod_{k=1}^{t} p\left(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1}, \boldsymbol{\Phi}\right) p\left(y_k | \boldsymbol{\theta}_k, \boldsymbol{\Phi}\right). \tag{9.1}$$

To apply MCMC using (9.1) within a particle filter, it would then be necessary, according to Fearnhead (2002), to store the entirety of particle state trajectories $\left\{\boldsymbol{\theta}_t^{(i)}\right\}_{i=1}^{N_p}$. As we've seen in Section 6.1 on page 69 this will have prohibitive computational costs for sequential inference as t evolves and would not constitute an online method. The concept of *sufficient statistics* then arises as a summary of the trajectories which MCMC moves will use. According to Fearnhead (2002), if we designate the current state and parameter

at time t as

$$\gamma = \{\boldsymbol{\theta}_t, \boldsymbol{\Phi}\},$$

and re-write the set

$$\begin{aligned}\{\boldsymbol{\theta}_{0:t}, \boldsymbol{\Phi}\} &= \{\{\boldsymbol{\theta}_t, \boldsymbol{\Phi}\}, \boldsymbol{\theta}_{0:t-1}\} \\ &= \{\gamma, \boldsymbol{\theta}_{0:t-1}\},\end{aligned}$$

we can denote the set

$$s = \{\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t\}$$

as a *sufficient statistic* if

$$p(\gamma|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t) = p(\gamma|s). \tag{9.2}$$

Still according to Fearnhead (2002), a necessary condition for sufficient statistics is that they can be factorised according to

$$p(\gamma, \boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t) = k_1[\gamma, s(\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t)]\, k_2[\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t]. \tag{9.3}$$

The main idea is that using the distribution $p(\gamma|s)$ is equivalent to using $p(\gamma|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t)$. Since we can factorise $p(\gamma, \boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t)$ as in (9.3), using (9.1) we can calculate $k_1[\gamma, s(\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t)]$ which in turn can be used to calculate $p(\gamma|s)$ up to a constant, that is

$$p(\gamma|s) \propto k_1[\gamma, s(\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t)].$$

Whereas in Fearnhead (2002) sufficient statistics are used as a summary of particle trajectories in order to update both states and parameters using MCMC moves, the Storvik filter, Storvik (2002) works by assuming that the parameter posterior $p(\boldsymbol{\Phi}|\boldsymbol{\theta}_{0:t}, \mathcal{D}_t)$ can be expressed as a function of a set of sufficient statistics s_t, such that

$$p(\boldsymbol{\Phi}|\boldsymbol{\theta}_{0:t}, \mathcal{D}_t) = p(\boldsymbol{\Phi}|s_t), \tag{9.4}$$

with an associated deterministic recursive update $\mathcal{S}(\cdot)$. For the purpose of estimating the parameter posterior, an assumption of sufficient statistics based filters is that $p(\boldsymbol{\Phi}|s_t)$ in (9.4) is analytically tractable, a constraint not present in the LW filter. In Storvik (2002), it is assumed that the posterior of $\boldsymbol{\Phi}$ will depend on a set of sufficient statistics s_t, themselves dependent of the state trajectory and data, that is

$$s_t = s_t(\boldsymbol{\theta}_{0:t}, \mathcal{D}_t).$$

If we consider the joint distribution of states and parameters, $p(\boldsymbol{\theta}_{0:t}, \boldsymbol{\Phi}|\mathcal{D}_t)$, we can factorise it as

$$\begin{aligned} p(\boldsymbol{\theta}_{0:t}, \boldsymbol{\Phi}|\mathcal{D}_t) &= \frac{p(\boldsymbol{\theta}_{0:t}, \boldsymbol{\Phi}, y_t|\mathcal{D}_{t-1})}{p(y_t|\mathcal{D}_{t-1})} \\ &= \frac{p(\boldsymbol{\theta}_{0:t-1}|\mathcal{D}_{t-1})\, p(\boldsymbol{\Phi}|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_{t-1})\, p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_{t-1}, \boldsymbol{\Phi})\, p(y_t|\boldsymbol{\theta}_{0:t}, \mathcal{D}_{t-1}, \boldsymbol{\Phi})}{p(y_t|\mathcal{D}_{t-1})}. \end{aligned}$$

By replacing the parameter's dependence on the states and data $\{\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_{t-1}\}$ by $s_{t-1} = s(\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_{t-1})$ as in (9.4) and using the Markovian property of the DGLM highlighted in (2.11), we have (up to a proportionality constant):

$$\begin{aligned} p(\boldsymbol{\theta}_{0:t}, \boldsymbol{\Phi}|\mathcal{D}_t) &= \frac{p(\boldsymbol{\theta}_{0:t-1}|\mathcal{D}_{t-1}) \overbrace{p(\boldsymbol{\Phi}|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_{t-1})}^{p(\boldsymbol{\Phi}|s_{t-1})} \overbrace{p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_{t-1}, \boldsymbol{\Phi})}^{p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi})} \overbrace{p(y_t|\boldsymbol{\theta}_{0:t}, \mathcal{D}_{t-1}, \boldsymbol{\Phi})}^{p(y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi})}}{p(y_t|\mathcal{D}_{t-1})} \\ &= \frac{p(\boldsymbol{\theta}_{0:t-1}|\mathcal{D}_{t-1})\, p(\boldsymbol{\Phi}|s_{t-1})\, p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi})\, p(y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi})}{p(y_t|\mathcal{D}_{t-1})} \\ &\propto p(\boldsymbol{\theta}_{0:t-1}|\mathcal{D}_{t-1})\, p(\boldsymbol{\Phi}|s_{t-1}) \underbrace{p(y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi})}_{\text{measurement}} \underbrace{p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi})}_{\text{system}}. \end{aligned}$$

Storvik (2002) proposes the use of the SIR framework, as described in Section 7.4 on page 91, to build the approximation of $p(\boldsymbol{\theta}_t|\mathcal{D}_t, \boldsymbol{\Phi})$ with an additional step to approximate $p(\boldsymbol{\Phi}|\boldsymbol{\theta}_{0:t}, \mathcal{D}_t)$ making use of (9.2). At each step t however, the sufficient statistics set s_t will need to update according to a deterministic function $\mathcal{S}(\cdot)$ such that

$$s_t = \mathcal{S}(s_{t-1}, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}, y_t). \tag{9.5}$$

It is important to note that the recursive deterministic update itself is sequential and online allowing for online estimation with the Storvik filter. In the next sections we will look at examples of sufficient statistics and their respective update for specific instances of DGLMs. By performing parameter estimation based on this set of sufficient statistics and separately from the state estimation, the Storvik filter aims at reducing particle impoverishment while reducing computational load due to the low-dimensionality of s_t. The recursive deterministic update of s_t will depend on the state and parameter estimates such as in (9.5).

The steps of the Storvik filter are summarised in Algorithm 9.1.

Although the Storvik filter works in a sample-resample framework, it can also be applied within a resample-sample framework.

Algorithm 9.1 Storvik's algorithm

initialisation

$\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0)$

$\boldsymbol{\Phi}_0 \sim p(\boldsymbol{\Phi})$

for $t \leftarrow 1$ to N_{obs}

for $i \leftarrow 1$ to N_p

sample $\boldsymbol{\Phi}_t^{(i)} \sim \pi_{\boldsymbol{\Phi}}\left(\boldsymbol{\Phi}|s_{t-1}^{(i)}\right)$

sample $\boldsymbol{\theta}_t^{(i)} \sim p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}^{(i)}, y_t, \boldsymbol{\Phi}_t^{(i)}\right)$

calculate weights

$$w_t^{(i)} = w_{t-1}^{(i)} \frac{p\left(\boldsymbol{\Phi}_t^{(i)}|s_{t-1}^{(i)}\right) p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}, \boldsymbol{\Phi}_t^{(i)}\right) p\left(y_t|\boldsymbol{\theta}_t^{(i)}, \boldsymbol{\Phi}_t^{(i)}\right)}{p\left(\boldsymbol{\Phi}_t^{(i)}|\boldsymbol{\theta}_{0:t-1}, \mathcal{D}_t\right) p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{0:t-1}^{(i)}, y_t, \boldsymbol{\Phi}_t^{(i)}\right)}$$

normalise weights

$$\widetilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{i=1}^{N_p} w_t^{(i)}}$$

calculate the *effective sample size*, $\widehat{ESS_t}$ as per (7.1).

if $\widehat{ESS_t} \leq N_{eff}$

resample according to $p(k(i) = l) = \tilde{w}_t^{(i)}$

for $i \leftarrow 1$ to N_p

set weights $\tilde{w}_t^{(i)} = \frac{1}{N_p}$

assign states $\boldsymbol{\theta}_t^{(i)} = \boldsymbol{\theta}_t^{\left(k^{(i)}\right)}$ and parameters $\boldsymbol{\Phi}^{(i)} = \boldsymbol{\Phi}^{\left(k^{(i)}\right)}$

for $i \leftarrow 1$ to N_p

update sufficient statistics

$$s_t^{(i)} = \mathcal{S}\left(s_{t-1}^{(i)}, \boldsymbol{\theta}_t^{(i)}, \boldsymbol{\theta}_{t-1}^{(i)}, y_t\right)$$

calculate state and parameter approximation

$$\begin{aligned} \hat{p}(\boldsymbol{\theta}_t|\mathcal{D}_t) &= \sum_{i=1}^{N_p} \tilde{w}_t^{(i)} \delta\left(\boldsymbol{\theta} - \boldsymbol{\theta}_t^{(i)}\right) \\ \hat{p}(\boldsymbol{\Phi}_t|\mathcal{D}_t) &= \sum_{i=1}^{N_p} \tilde{w}_t^{(i)} \delta\left(\boldsymbol{\Phi} - \boldsymbol{\Phi}_t^{(i)}\right) \end{aligned}$$

9.1 Sufficient statistics

9.1.1 Normal DLM

To calculate the sufficient statistics of a Gaussian DLM in the form of Section 2.3.1 we start by writing the joint posterior of the parameters conditional on the data and states, that is:

$$\pi\left(\boldsymbol{\Phi}|\boldsymbol{\theta}_{0:t}, \mathcal{D}_t\right).$$

Using the same assumptions as in West & Harrison (1997), we take independent priors for $\boldsymbol{\Phi} = \{V, \mathsf{W}\}$, that is

$$\pi\left(V, \mathsf{W}|\mathcal{D}_0\right) = \pi\left(V|\mathcal{D}_0\right)\pi\left(\mathsf{W}|\mathcal{D}_0\right)$$

with

$$\begin{aligned}\pi\left(V|\mathcal{D}_0\right) &= \mathcal{IG}\left(\alpha_0^V, \beta_0^V\right)\\ \pi\left(\mathsf{W}|\mathcal{D}_0\right) &= \mathcal{IW}\left(\alpha_0^W, \beta_0^W\right)\end{aligned}$$

The full conditional update of V will then be

$$\begin{aligned}\pi\left(V|\mathcal{D}_t, \boldsymbol{\theta}_{0:t}\right) &\propto \mathcal{N}\left(y_t|\mathsf{F}^T\boldsymbol{\theta}_t, V\right)\mathcal{IG}\left(V|\alpha_{t-1}^V, \beta_{t-1}^V\right)\\ &\propto \underbrace{\frac{1}{\sqrt{V}}\exp\left\{-\frac{\left(y_t - \mathsf{F}^T\boldsymbol{\theta}_t\right)^2}{2V}\right\}}_{\mathcal{N}\left(y_t|\mathsf{F}^T\boldsymbol{\theta}_t, \sigma^2\right)} \times \underbrace{V^{-\alpha_{t-1}^V-1}\exp\left\{-\frac{\beta_{t-1}^V}{V}\right\}}_{\mathcal{IG}\left(\sigma^2|\alpha_{t-1}^V, \beta_{t-1}^V\right)}\\ &\propto V^{-\frac{1}{2}-\alpha_{t-1}^V-1}\exp\left\{-\frac{\beta_{t-1}^V}{V} + \frac{\left(y_t - \mathsf{F}^T\boldsymbol{\theta}_t\right)^2}{2V}\right\}\\ &\propto V^{\underbrace{-\left(\frac{1}{2}+\alpha_{t-1}^V\right)}_{\alpha_t^V}-1}\exp\left\{-\frac{\underbrace{\beta_{t-1}^V + \frac{\left(y_t - \mathsf{F}^T\boldsymbol{\theta}_t\right)^2}{2}}_{\beta_t^V}}{V}\right\}\\ &\propto \mathcal{IG}\left(\frac{1}{2} + \alpha_{t-1}^V, \beta_{t-1}^V + \frac{1}{2}\left(y_t - \mathsf{F}^T\boldsymbol{\theta}_t\right)^2\right).\end{aligned}$$

It follows that

$$V|\boldsymbol{\theta}_{0:t}, \mathcal{D}_t \sim \mathcal{IG}\left(\alpha_0^V + \frac{1}{2}t, \beta_0^V + \frac{1}{2}\sum_{k=1}^{t}\left(y_k - \mathsf{F}^T\boldsymbol{\theta}_k\right)^2\right),$$

and taking into account the sufficient statistics condition in (9.2) we can see that

$$\begin{aligned} p\left(V|\boldsymbol{\theta}_{0:t}, \mathcal{D}_t\right) &= \mathcal{IG}\left(\alpha_0^V + \frac{1}{2}\underbrace{t}_{s_1^V}, \beta_0^V + \frac{1}{2}\underbrace{\sum_{k=1}^{t}\left(y_k - \mathsf{F}^T\boldsymbol{\theta}_k\right)^2}_{s_2^V}\right) \\ &= p\left(V|s\right) = \mathcal{IG}\left(\alpha_0^V + \frac{s_1^V}{2}, \beta_0^V + \frac{1}{2}s_2^V\right), \end{aligned}$$

so that the sufficient statistics for V will be

$$\begin{aligned} s_{1,t}^V &= 1 \\ s_{2,t}^V &= \left(y_t - \mathsf{F}^T\boldsymbol{\theta}_t\right)^2, \end{aligned}$$

with a recursive update $\mathcal{S}\left(\cdot\right)$ in the form

$$\begin{aligned} s_{1,t}^V &= s_{1,t-1}^V + 1 \\ s_{2,t}^V &= s_{2,t-1}^V + \left(y_t - \mathsf{F}^T\boldsymbol{\theta}_t\right)^2. \end{aligned}$$

The full conditional update of W can be obtained in an analogous way, that is:

$$\begin{aligned} \pi\left(\mathsf{W}|\mathcal{D}_t, \boldsymbol{\theta}_{0:t}\right) &\propto \mathcal{N}\left(\boldsymbol{\theta}_t|\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}\right)\mathcal{IW}\left(\mathsf{W}_t|\alpha_{\mathsf{W},t-1}, \Psi_{\mathsf{W},t-1}^{-1}\right) \\ &\propto \left(2\pi\right)^{-p/2}\left|\mathsf{W}\right|^{-1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right)^T\mathsf{W}^{-1}\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right)\right\} \\ &\times \frac{\left|\Psi_{\mathsf{W},t-1}\right|^{\alpha_{\mathsf{W},t-1}}}{2^{\frac{\alpha_{\mathsf{W},t-1}p}{2}}\Gamma_p\left(\frac{\alpha_{\mathsf{W},t-1}}{2}\right)}\left|\mathsf{W}\right|^{-\frac{\alpha_{\mathsf{W},t-1}+p+1}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\Psi_{W,t-1}\mathsf{W}^{-1}\right)\right\} \\ &\propto \left|\mathsf{W}\right|^{-1/2}\left|\mathsf{W}\right|^{-\frac{\alpha_{\mathsf{W},t-1}+p+1}{2}} \\ &\times \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right)\mathsf{W}^{-1}\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right)^T + \mathrm{tr}\left(\Psi_{\mathsf{W},t-1}\mathsf{W}^{-1}\right)\right]\right\} \\ &\propto \left|\mathsf{W}\right|^{\frac{-\alpha_{\mathsf{w},t-1}-1-p-1}{2}} \times \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right)\mathsf{W}^{-1}\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right)^T + \Psi_{\mathsf{W},t-1}\mathsf{W}^{-1}\right)\right\}, \end{aligned}$$

which can be further simplified into

$$\pi\left(\mathsf{W}|\mathcal{D}_t,\boldsymbol{\theta}_{0:t}\right)\propto|\mathsf{W}|^{-\overbrace{\frac{\alpha_{\mathsf{W},t-1}+1}{2}}^{\alpha_{\mathsf{W},t}}+p+1}\times\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\underbrace{\left[\left(\boldsymbol{\theta}_t-\mathsf{G}\boldsymbol{\theta}_{t-1}\right)\left(\boldsymbol{\theta}_t-\mathsf{G}\boldsymbol{\theta}_{t-1}\right)^T+\Psi_{\mathsf{W},t-1}\right]}_{\Psi_{\mathsf{w},t}}\mathsf{W}^{-1}\right)\right\},$$

resulting in

$$\pi\left(\mathsf{W}|\mathcal{D}_t,\boldsymbol{\theta}_{0:t}\right)\propto\mathcal{IW}\left(\alpha_{\mathsf{W},t-1}+1,\left(\boldsymbol{\theta}_t-\mathsf{G}\boldsymbol{\theta}_{t-1}\right)\left(\boldsymbol{\theta}_t-\mathsf{G}\boldsymbol{\theta}_{t-1}\right)^T+\Psi_{\mathsf{W},t-1}\right). \tag{9.6}$$

where Ψ is a scale matrix and α_{W} represents the degrees of freedom. As with the calculations for V_t, since

$$\begin{aligned}\mathsf{W}|\boldsymbol{\theta}_{0:t},\mathcal{D}_t&\sim\mathcal{IW}\left(\alpha_0^{\mathsf{W}}+\underbrace{t}_{s_1^W},\beta_0^V+\underbrace{\sum_{k=1}^{t}\left(\boldsymbol{\theta}_k-\mathsf{G}\boldsymbol{\theta}_{k-1}\right)\left(\boldsymbol{\theta}_k-\mathsf{G}\boldsymbol{\theta}_{k-1}\right)^T}_{s_2^W}\right)\\&\sim\mathcal{IW}\left(\alpha_0^{\mathsf{W}}+s_1^W,\beta_0^V+s_2^W\right),\end{aligned}$$

from which we extract the sufficient statistics

$$\begin{aligned}s_{1,t}^W&=1\\s_{2,t}^W&=\left(\boldsymbol{\theta}_t-\mathsf{G}\boldsymbol{\theta}_{t-1}\right)\left(\boldsymbol{\theta}_t-\mathsf{G}\boldsymbol{\theta}_{t-1}\right)^T\end{aligned}$$

with the recursive update $\mathcal{S}\left(\cdot\right)$

$$\begin{aligned}s_{1,t}^W&=s_{1,t-1}^W+1\\s_{2,t}^W&=s_{2,t-1}^W+\left(\boldsymbol{\theta}_t-\mathsf{G}\boldsymbol{\theta}_{t-1}\right)\left(\boldsymbol{\theta}_t-\mathsf{G}\boldsymbol{\theta}_{t-1}\right)^T.\end{aligned}$$

9.1.2 Binomial and Poisson DLM

In the Binomial DLM case, as described in Section 2.3.3, and the PoDLM, Section 2.3.2, we can see that we do not have parameters to be estimated in the observational model (2.25) and (2.22), resulting in $\boldsymbol{\Phi}=\{\mathsf{W}\}$. In this case, the sufficient statistics and respective recursions will be analogous to the ones derived for $\pi\left(\mathsf{W}|\boldsymbol{\theta}_{0:t},\mathcal{D}_t\right)$ in the NDLM in Section 9.1.1.

Example. Storvik filter for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM.

In this case we will apply the Storvik filter to the model and data presented in the LW example on page 112. For a locally constant model, the sufficient statistics structure will simplify, since in this case we have

$$\begin{aligned} p\left(\sigma^2|\theta_{0:t}, \mathcal{D}_t\right) &= \mathcal{IG}\left(\alpha_0^{\sigma^2} + \frac{1}{2}t, \beta_0^{\sigma^2} + \frac{1}{2}\sum_{k=1}^{t}(y_k - \theta_k)^2\right) \\ p\left(\tau^2|\theta_{0:t}, \mathcal{D}_t\right) &= \mathcal{IG}\left(\alpha_0^{\tau^2} + \frac{1}{2}t, \beta_0^{\tau^2} + \frac{1}{2}\sum_{k=1}^{t}(\theta_k - \theta_{k-1})^2\right), \end{aligned}$$

it follows that the sufficient statistics set will be

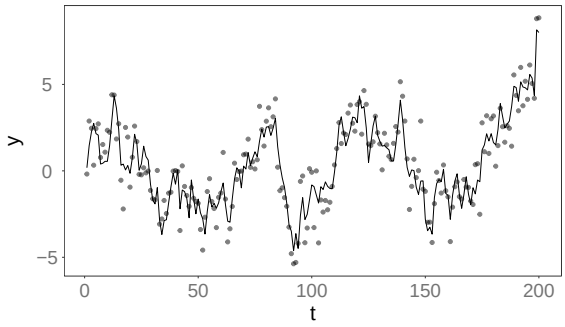
$$s_t = \left\{1, (y_t - \theta_t)^2, (\theta_t - \theta_{t-1})^2\right\}$$

with a recursive update

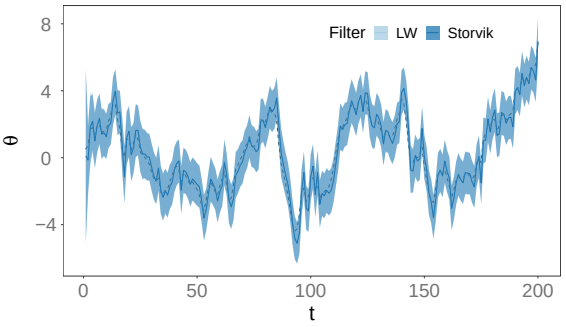
$$\begin{aligned} s_{1,t} &= s_{1,t-1} + 1 \\ s_{2,t} &= s_{2,t-1} + (y_t - \theta_t)^2 \\ s_{3,t} &= s_{3,t-1} + (\theta_t - \theta_{t-1})^2. \end{aligned}$$

As in the LW example, the Storvik filter is the fully adapted version, using the optimal importance density, since we are dealing with a NDLM.

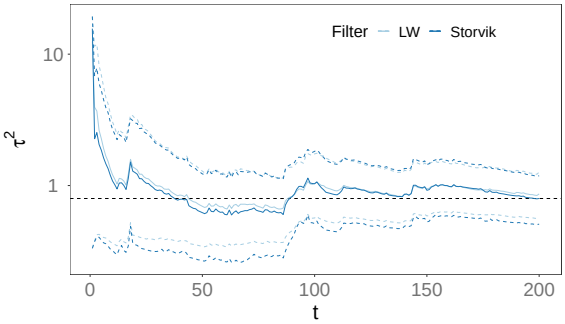
This estimation was performed using $N_p = 10^5$ particles and the resampling scheme used was Multinomial with a static checkpoint of $n = 1$. The prior used for the parameter set $\boldsymbol{\Phi}$ were $\tau_0^2 \sim \mathcal{IG}(1,1)$ and $\nu_0^2 \sim \mathcal{IG}(1,1)$. Regarding the state estimation (Figure 9.1b) the results are comparable between Storvik and LW. When considering parameter estimation, we can see (Figures 9.1e and 9.1f) that the Storvik filter and LW parameter and state estimates are consistent, which is expected for a simple model with a small number of observations and especially when using the optimal importance density. The posterior mean (and standard deviation) at time $t = N_{obs}$ are, respectively for Storvik and LW, $\overline{\tau}^2 = 0.7921, \sigma_{\tau^2} = 0.2116, \overline{\nu}^2 = 1.1136, \sigma_{\nu^2} = 0.2092$ and $\overline{\tau}^2 = 0.8402, \sigma_{\tau^2} = 0.2034, \overline{\nu}^2 = 1.1253, \sigma_{\nu,^2} = 0.2032$, compared to "true" values of $\overline{\tau}^2 = 0.7932, \sigma_{\tau^2} = 0.2390$ and $\overline{\nu}^2 = 1.2011, \sigma_{\nu^2} = 0.2263$.



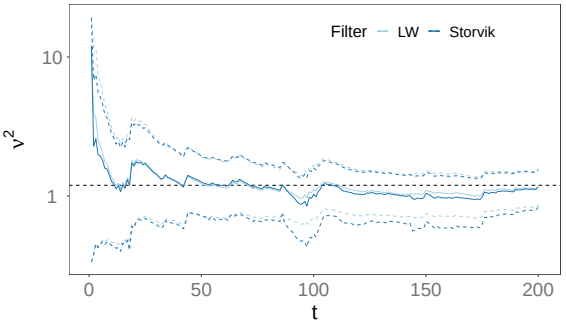
(a) True state (*line*) and data (*points*) for a simulated $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM



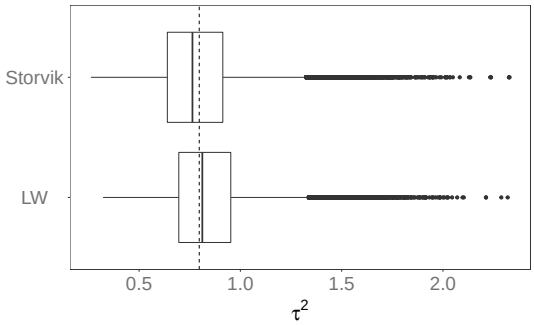
(b) Mean of the state posterior estimation with 90% equitailed credibility intervals. Dashed line represents true state.



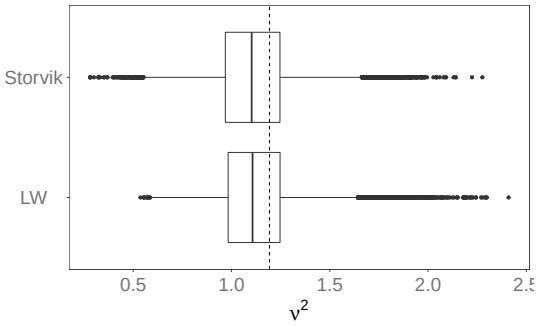
(c) τ^2 posterior estimation history. Coloured line is posterior mean and coloured dashed lines are 90% equitailed credibility intervals. Horizontal line represents the truth (*log scale*).



(d) ν^2 posterior estimation history. Coloured line is posterior mean and coloured dashed lines are 90% equitailed credibility intervals. Horizontal line represents the truth (*log scale*).



(e) $p\left(\tau^2|\mathcal{D}_t\right)$ posterior using Storvik and L&W at $t = 200$. Vertical line represents truth.



(f) $p\left(\nu^2|\mathcal{D}_t\right)$ posterior using Storvik and L&W at $t = 200$. Vertical line represents truth.

Figure 9.1: State and parameter estimation for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM using LW and Storvik

Chapter 10

Particle Learning

Particle Learning (PL), first introduced in Carvalho *et al.* (2010), similarly to the Storvik filter relies on the concept of sufficient statistics, denoted as the *essential state vector* z_t, a more general sufficient statistic set which can include not only parameter sufficient statics to marginalise the parameters, as in (9.4), but also other quantities that could summarise the state and parameter approximation. As in Storvik, we assume that we can devise a sufficient statistic structure for the parameters, such as

$$\pi\left(\boldsymbol{\Phi}|\boldsymbol{\theta}_{0:t}, \mathcal{D}_t\right) = \pi\left(\boldsymbol{\Phi}|s_t^{\boldsymbol{\Phi}}\right), \tag{10.1}$$

where $s_t^{\boldsymbol{\Phi}}$ is the set of *parameter sufficient statistics*, with a recursive deterministic update, such that

$$s_t^{\boldsymbol{\Phi}} = \mathcal{S}\left(s_{t-1}^{\boldsymbol{\Phi}}, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}, y_t\right).$$

PL tries to perform estimation in a *resample-sample* framework (related to the APF in Section 7.6 on page 96) using *full-adaption* with the usage, as mentioned previously, of an additional set of *state sufficient statistics*, $s_t^{\boldsymbol{\theta}}$ whenever these are analytically available. These quantities are associated with a recursive deterministic update $\mathcal{K}(\cdot)$ similar to the Kalman filter recursions presented in Section 4, such that

$$s_t^{\boldsymbol{\theta}} = \mathcal{K}\left(s_{t-1}^{\boldsymbol{\theta}}, \boldsymbol{\theta}_{t-1}, y_t\right).$$

The set of quantities comprised of the state and parameter sufficient statistics defines the *essential state vector* z_t, such that

$$z_t = \left\{s_t^{\boldsymbol{\Phi}}, s_t^{\boldsymbol{\theta}}, \boldsymbol{\theta}_t, \boldsymbol{\Phi}\right\}.$$

In PL the main idea is to pre-select the essential state vector particles, $\left\{z_{t-1}^{(i)}\right\}_{i=1}^{N_p}$ using an auxiliary weight $\lambda_t^{(i)}$ based on a predictive density. Similarly to Storvik, the state particles $\boldsymbol{\theta}_t^{(i)}$ could then be sampled from an importance density conditioned on the selected $\left\{z_{t-1}^{(k)}\right\}_{k=1}^{N_p}$. The essential state vector would then be propagated according the deterministic updates and finally the parameters could also be simulated from a proposal conditioned on the essential vector.

According to Carvalho *et al.* (2010) if at time $t-1$ and in possession of observation y_{t-1} we have a discrete approximation to $p\left(z_{t-1}|\mathcal{D}_{t-1}\right)$ with $\left\{z_{t-1}^{(i)}\right\}_{i=1}^{N_p}$ such that

$$p\left(z_{t-1}|\mathcal{D}_{t-1}\right) \approx \frac{1}{N_P}\sum_{i=1}^{N_p}\delta\left(z_{t-1}-z_{t-1}^{(i)}\right). \tag{10.2}$$

The objective is then to use $p\left(z_{t-1}|\mathcal{D}_{t-1}\right)$ to calculate $p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right)$ once the new observation y_t arrives. According to Carvalho *et al.* (2010), the essential state vector is updated following

$$\begin{aligned}
p\left(z_{t-1}|\mathcal{D}_t\right) &= p\left(z_{t-1}|\mathcal{D}_{t-1},y_t\right) \\
&= \frac{p\left(y_t|z_{t-1},\mathcal{D}_{t-1}\right)p\left(z_{t-1}|\mathcal{D}_{t-1}\right)}{p\left(y_t|\mathcal{D}_{t-1}\right)} \\
&= \frac{p\left(y_t|z_{t-1}\right)p\left(z_{t-1}|\mathcal{D}_{t-1}\right)}{p\left(y_t|\mathcal{D}_{t-1}\right)} \\
&\propto p\left(y_t|z_{t-1}\right)p\left(z_{t-1}|\mathcal{D}_{t-1}\right) \qquad (10.3)\\
p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right) &= \int p\left(\boldsymbol{\theta}_t|z_{t-1},\mathcal{D}_t\right)p\left(z_{t-1}|\mathcal{D}_t\right)dz_{t-1}. \qquad (10.4)
\end{aligned}$$

According to Carvalho *et al.* (2010), if we assume that a discrete approximation to the essential state vector prior (10.3) is

$$p\left(z_{t-1}|\mathcal{D}_t\right) \approx \sum_{t=1}^{N_p}\lambda_t^{(i)}\delta\left(z_{t-1}-z_{t-1}^{(i)}\right),$$

with normalised weights

$$\lambda_t^{(i)} = \frac{p\left(y_t|z_{t-1}^{(i)}\right)}{\sum_{j=1}^{N_p}p\left(y_t|z_{t-1}^{(j)}\right)},$$

this approximation of $p\left(z_{t-1}|\mathcal{D}_t\right)$ can then be used in conjunction with (10.4) to sample the current states according to

$$\boldsymbol{\theta}_t^{(i)} \sim p\left(\boldsymbol{\theta}_t|z_{t-1}^{(i)},\mathcal{D}_t\right).$$

Carvalho *et al.* (2010) notes that the deterministic state-sufficient-statistics recursion, $\mathcal{K}(\cdot)$, can then be applied to the samples $\left\{\boldsymbol{\theta}_t^{(i)}\right\}_{i=1}^{N_p}$ to generate the new state sufficient statistics as

$$s_t^{\boldsymbol{\theta}} = \mathcal{K}\left(s_{t-1}^{\boldsymbol{\theta}}, \boldsymbol{\theta}_t, \Phi, y_t\right).$$

The steps to update the parameters will be analogous to the updates for the parameter sufficient statistics of Storvik, as detailed in Chapter 9 on page 116. To calculate the posterior $p(z_t|\mathcal{D}_t)$, we start by expressing it as

$$p(z_t|\mathcal{D}_t) = \int p(z_t|z_{t-1}, \mathcal{D}_t)\, p(z_{t-1}|\mathcal{D}_t)\, dz_{t-1}.$$

Carvalho *et al.* (2010) summarises the *fully adapted* case as obtaining samples from $p(\boldsymbol{\theta}_t, z_{t-1}|\mathcal{D}_t)$ with respect to draws from $p(z_{t-1}|\mathcal{D}_t)\, p(\boldsymbol{\theta}_t|z_{t-1}, \mathcal{D}_t)$ with importance weights

$$w_t \propto \frac{p(\boldsymbol{\theta}_t, z_t|\mathcal{D}_t)}{p(z_{t-1}|\mathcal{D}_t)\, p(\theta_t|z_{t-1}, \mathcal{D}_t)} = 1.$$

Assuming that a state-sufficient-statistic structure can be devised, we can state that

$$p(\boldsymbol{\theta}_t|\mathcal{D}_t) = \int p(\boldsymbol{\theta}_t|z_t)\, p(z_t|\mathcal{D}_t)\, dz_t,$$

which implies the filtering recursion

$$p(z_t|\mathcal{D}_t) = \int p(z_t|z_{t-1}, \boldsymbol{\theta}_t, y_t)\, p(z_{t-1}, \boldsymbol{\theta}_t|\mathcal{D}_t)\, dz_{t-1} d\boldsymbol{\theta}_t. \tag{10.5}$$

Since the joint density $p(\boldsymbol{\theta}_t, z_{t-1}|\mathcal{D}_t)$ can be factorised as

$$p(\boldsymbol{\theta}_t, z_{t-1}|\mathcal{D}_t) \propto p(y_t|z_{t-1})\, p(\boldsymbol{\theta}_t|z_{t-1}, y_t)\, p(z_{t-1}|\mathcal{D}_{t-1}), \tag{10.6}$$

assuming we have a $t-1$ approximation of $p(\boldsymbol{\theta}_{t-1}|z_{t-1})$, that is, we can calculate

$$p(\boldsymbol{\theta}_t|z_{t-1}, y_t) = \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, y_t)\, p(\boldsymbol{\theta}_{t-1}|z_{t-1})\, d\boldsymbol{\theta}_{t-1},$$

then in this case we could marginalise not only the parameters through the parameter sufficient statistics (as in Storvik) but also the states according the state sufficient statistics, that is we would only need to track the conditional sufficient statistics (Carvalho *et al.* (2010)).

Although the marginalisation of states and parameters (in the fully adapted case) helps to mitigate weight degeneracy and particle impoverishment, as discussed in Chopin *et al.*

(2010), the problem is not completely solved. Although the state and parameter particles can be replenished from the essential state vector z_t, $p\left(z_t|\mathcal{D}_t\right)$ is defined in terms of a discrete approximation as shown in (10.2). This approximation, coupled with the fact that we are applying resampling at each time step t according to the *resample-sample* framework, will lead to a degeneracy of s_t itself (Chopin *et al.* (2010)), which will be inevitable unless the number of particles N_p increases exponentially with t (Chopin *et al.* (2010); Del Moral *et al.* (2006)).

In the next sections we will look at the details of a fully adapted PL case (the NDLM) and also the PL case in non-linear models and a general algorithm for the PL is presented in Algorithm 10.1.

Algorithm 10.1 Particle Learning

initialisation

$\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0)$

$\boldsymbol{\Phi}_0 \sim p(\boldsymbol{\Phi})$

$\{z_0\}_{i=1}^{N_p} = \left\{s_0^{\boldsymbol{\Phi}}, s_0^{\boldsymbol{\theta}}, \boldsymbol{\theta}_0, \boldsymbol{\Phi}_{\mathbf{0}}\right\}_{i=1}^{N_p}$

for $t \leftarrow 1$ to N_{obs}

for $i \leftarrow 1$ to N_p

calculate auxiliary weights

$$\lambda_t^{(i)} \propto p\left(y_t | z_{t-1}^{(i)}\right)$$

resample $\left\{z_{t-1}^{(i)}\right\}_{i=1}^{N_p}$ according the auxiliary weights $\left\{\lambda_t^{(i)}\right\}_{i=1}^{N_p}$

for $i \leftarrow 1$ to N_p

sample $\boldsymbol{\theta}_t^{(i)} \sim p\left(\boldsymbol{\theta}_t | z_{t-1}^{\left(k^{(i)}\right)}, y_t\right)$

propagate state sufficient statistics according to

$$s_t^{\boldsymbol{\theta}^{(i)}} = \mathcal{K}\left(s_{t-1}^{\boldsymbol{\theta}^{\left(k^{(i)}\right)}}, \boldsymbol{\theta}_t^{(i)}, y_t\right)$$

propagate parameter sufficient statistics according to

$$s_t^{\boldsymbol{\Phi}^{(i)}} = \mathcal{S}\left(s_{t-1}^{\boldsymbol{\Phi}^{\left(k^{(i)}\right)}}, \boldsymbol{\theta}_t^{(i)}, y_t\right)$$

sample parameters according to

$$\boldsymbol{\Phi}^{(i)} \sim p\left(\boldsymbol{\Phi} | s_t^{\boldsymbol{\Phi}^{(i)}}\right)$$

set $z_t^{(i)} = \left\{s_t^{\boldsymbol{\Phi}^{(i)}}, s_t^{\boldsymbol{\theta}^{(i)}}, \boldsymbol{\theta}_t^{(i)}, \boldsymbol{\Phi}^{(i)}\right\}$

10.1 Normal DLM

For the Normal DLM, the implementation of the PL is straightforward due to the fact that the *predictive density* is readily available, as described in (4.9) in the Kalman filter Section 4. Considering (10.6)

$$p\left(s_{t-1}^{\boldsymbol{\theta}}, \boldsymbol{\theta}_t | \mathcal{D}_t\right) \propto p\left(y_t | s_{t-1}^{\boldsymbol{\theta}}\right) p\left(\boldsymbol{\theta}_t | s_{t-1}^{\boldsymbol{\theta}}, y_t\right) p\left(s_{t-1}^{\boldsymbol{\theta}} | \mathcal{D}_{t-1}\right),$$

we can see that we are able to calculate, in the NDLM case, the densities $p\left(y_t|s_{t-1}^{\boldsymbol{\theta}}\right)$, $p\left(\boldsymbol{\theta}_t|s_{t-1}^{\boldsymbol{\theta}}, y_t\right)$ as well as the deterministic updated $\mathcal{K}(\cdot)$ in a closed form. This allows us to write PL for the NDLM as a *fully adapted* version. If we consider a *state sufficient statistic* set as comprising the first and second KF moments at time $t-1$, such that

$$s_{t-1}^{\boldsymbol{\theta}} = \{\mathbf{m}_{t-1}, \mathsf{C}_{t-1}\},$$

for a Normal DLM in the standard form of Section 2.3.1, we then have the predictive density, given the state sufficient statistics as

$$\begin{aligned} p\left(y_t|s_{t-1}^{\boldsymbol{\theta}}\right) &= p\left(y_t|\mathbf{m}_{t-1}, \mathsf{C}_{t-1}\right) \\ &= \mathcal{N}\left(\underbrace{\mathsf{F}^T\mathsf{G}\mathbf{m}_{t-1}}_{f_t}, \underbrace{\mathsf{F}^T\left[\mathsf{G}\mathsf{C}_{t-1}\mathsf{G}^T + \mathsf{W}\right]\mathsf{F} + V}_{Q_t}\right). \end{aligned}$$

Given the predictive density, we can then resample the vector $\left\{z_{t-1}^{(i)}\right\}_{i=1}^{N_p}$ based on the auxiliary weights

$$\lambda_t^{(i)} \propto p\left(y_t|s_{t-1}^{\boldsymbol{\theta}}\right).$$

After resampling, the state particles can then be sampled from

$$p\left(\boldsymbol{\theta}_t|s_{t-1}^{\boldsymbol{\theta}}, y_t\right) = \mathcal{N}\left(\left(\mathbf{I} - \mathbf{K}_t\mathsf{F}^T\right)\boldsymbol{\theta}_{t-1} + \mathbf{K}_t y_t, \left(\mathbf{I} - \mathbf{K}_t\mathsf{F}^T\right)\mathsf{W}\right)$$

The state sufficient statistics update, for each particle i, will then consist of the KF filter recursions as presented in (4.11) and (4.12), that is

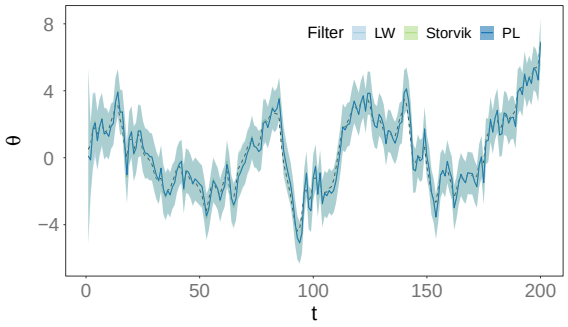
$$\begin{aligned} s_t^{\boldsymbol{\theta}} &= \mathcal{K}\left(s_{t-1}^{\boldsymbol{\theta}}, \boldsymbol{\theta}_t, y_t\right) \\ &= \left\{\mathbf{a}_t + \mathsf{R}_t\mathsf{F}\left(\mathsf{F}^T\mathsf{R}_t\mathsf{F} + V\right)^{-1}(y - f_t), \mathsf{R}_t - \mathsf{R}_t\mathsf{F}\left(\mathsf{F}^T\mathsf{R}_t\mathsf{F} + \mathsf{W}\right)^{-1}\mathsf{F}\mathsf{R}_t\right\}, \end{aligned}$$

with

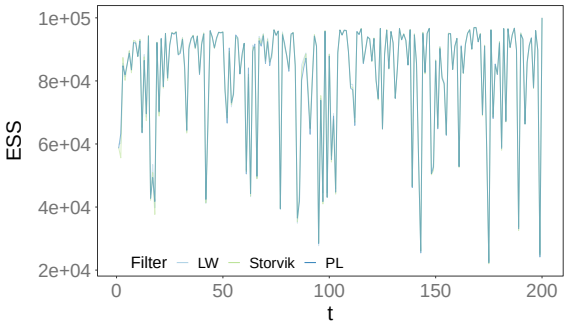
$$\begin{aligned} \mathbf{a}_t &= \mathsf{G}\mathbf{m}_{t-1} \\ \mathsf{R}_t &= \mathsf{G}\mathsf{C}_{t-1}\mathsf{G}^T + \mathsf{W}. \end{aligned}$$

The parameters sufficient statistic update and sampling will take the same form as in the Storvik filter, as detailed in Section 9.1 on page 120.

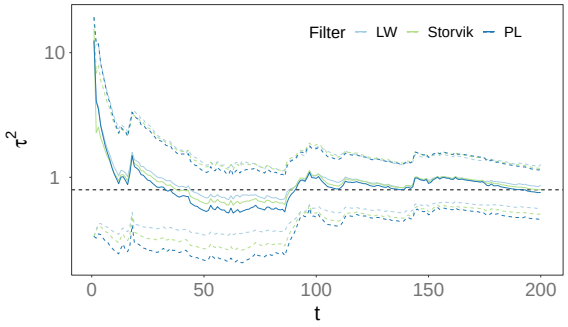
Example. PL for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM.



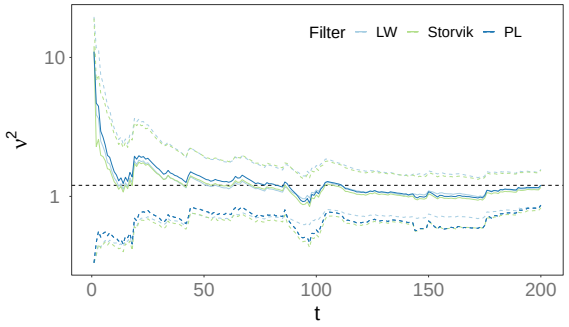
(a) $\hat{\theta}_{0:t}$, colour lines represent posterior mean and shaded area the 90% equitailed credibility interval.



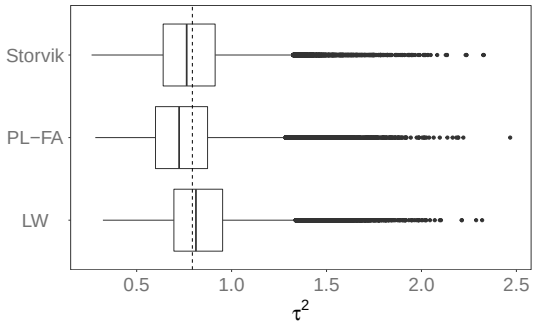
(b) $\widehat{ESS}$



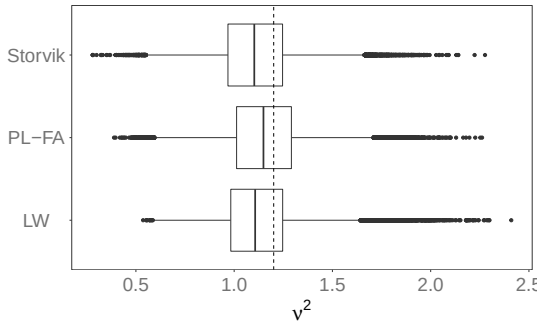
(c) τ^2 posterior estimation history. Solid line represents the posterior mean and dashed lines the 90% equitailed credibility interval.



(d) ν^2 posterior estimation history. Solid lines represent the posterior mean and dashed lines the 90% equitailed credibility interval.



(e) $p\left(\tau^2|\mathcal{D}_t\right)$ marginal at $t = 200$



(f) $p\left(\nu^2|\mathcal{D}_t\right)$ marginal at $t = 200$

Figure 10.1: State and parameter estimation using fully adapted LW, Storvik and PL for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with $\boldsymbol{\Phi} = \left\{\tau^2, \nu^2\right\} = \{0.75, 1.25\}$

Here we consider a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM as the one presented for the LW in the example 8 on page 112. The parameter set used was $\boldsymbol{\Phi} = \left\{\tau^2, \nu^2\right\} = \{0.75, 1.25\}$. The initial parameter priors for all filters were

$$\tau_0^2 \sim \mathcal{IG}(1,1), \qquad \nu_0^2 \sim \mathcal{IG}(1,1).$$

The PL algorithm was implemented in the fully adapted variant, that is, following the specification in Section 10.1 and the number of particles was the same for all filters, $N_p = 10^5$ and using stratified resampling. The state posterior estimation can be viewed in Figure 10.1a and the $\widehat{ESS}$ in Figure 10.1b. The parameter posterior estimation history for τ^2 and ν^2 can be viewed in Figures 10.1c and 10.1d respectively and the parameters posterior at time $t = N_{obs}$ can be viewed in Figures 10.1e and 10.1f. The posterior mean of τ^2 at $t = N_{obs}$ was, respectively for Storvik, PL-FA and L&W, $0.7921, 0.7526$ and 0.8402 and for ν^2 $1.1136, 1.1553$ and 1.1253. The "true" posterior mean is $\overline{\tau}^2 = 0.7932$ and $\overline{\nu}^2 = 1.2011$.

10.2 Non-linear DGLMs

In non-linear DGLMs it might not be possible to directly apply the *fully adapted* formulation, due to the inability to specify a set of state sufficient statistics and directly evaluate a predictive density $p\left(y_t|z_{t-1}\right)$. However, according to Nemeth *et al.* (2014), we can use the advantages of the APF *resample-sample* along with the parameter sufficient statistics structure (which we know are available for the general DGLMs).

If we assume that at time t we have the state and parameter approximation

$$\left\{\boldsymbol{\theta}_{t-1}^{(i)}, \boldsymbol{\Phi}_{t-1}^{(i)}\right\}_{i=1}^{N_p},$$

along with the sufficient statistics $\left\{s_{t-1}^{(i)}\right\}_{i=1}^{N_p}$ and importance weights $\left\{w_{t-1}^{(i)}\right\}_{i=1}^{N_p}$, then as in the APF case in Section 7.6, we can pre-select the particles using the auxiliary weight

$$\lambda_t^{(i)} \propto p\left(y_t|\boldsymbol{\mu}_t^{(i)}\right) w_{t-1}^{(i)},$$

where $\boldsymbol{\mu}_t^{(i)}$ is a characterisation of $p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(i)}, \boldsymbol{\Phi}_{t-1}^{(i)}\right)$ such that

$$\boldsymbol{\mu}_t^{(i)} = \mathrm{E}\left[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(i)}, \boldsymbol{\Phi}_{t-1}^{(i)}\right].$$

It is clear that in this case we will not be able to marginalise the states using state sufficient statistics, however after the resampling step the remaining step of PL can still be applied. In this case however, since full adaption is not available, exact *i.i.d.* samples (such as described in Section 6.2.1 on page 73) will not be produced as the auxiliary weights $\lambda_t^{(i)}$

must be corrected as in the APF such that

$$w_t^{(i)} \propto \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)^{(k)}}, \boldsymbol{\Phi}_t^{(i)^{(k)}}\right)}{p\left(y_t|\boldsymbol{\mu}_t^{(i)}, \boldsymbol{\Phi}^{(i)}\right)}.$$

A general algorithm for PL in non-linear DGLMs is presented in Algorithm 10.2.

Example. PL for a $\mathcal{M} = \{\mathcal{P}(1)\}$ PoDLM.

Here we consider the results of state and parameter estimation for an arbitrarily small dataset ($N_{obs} = 200$) from a realisation of a $\mathcal{M} = \{\mathcal{P}(1)\}$ PoDLM with a randomly chosen parameter $\boldsymbol{\Phi} = \{\tau^2\} = 0.2$. The state priors are set according to $\theta_0 \sim \mathcal{N}(0, 1000)$ and the parameter prior as $\tau_0^2 \sim \mathcal{IG}(1, 1)$. A comparison is made using the Storvik and PL filters, both with $N_p = 10^5$. The importance density used was the prior, $p(\theta_t|\theta_{t-1}, y_t) = \mathcal{N}\left(\theta_{t-1}, \tau^2\right)$ and the parameter sufficient statistics and recursive update were defined as in Section 9. Stratified resampling was used with a static checkpoint of $n = 1$. In Figures 10.2a and 10.2b we can see the estimation of $\theta_{0:t}$ using the Storvik and PL filter, respectively. Dashed line represents the mean of "true" state[1]. In Figure 10.2c we can see the estimation history of τ^2 using Storvik and PL and in Figure 10.2d we can see the marginal the of τ^2 at $t = N_{obs}$ using Storvik and PL with the vertical dashed line representing the value of τ^2 used for the data simulation. From Table 10.1 we can see that the results of Storvik and PL are consistent both in terms of parameter posterior and state MSE (comparing the filter's state posterior mean with the PMMH state posterior mean).

Method	$\overline{\tau}^2$	MSE_θ
Storvik	0.2243 (0.0325)	0.0816
PL	0.2261 (0.0329)	0.0815
PMMH	0.2162 (0.0316)	—

Table 10.1: Summary of parameter posterior mean and standard deviation (in brackets) at time $t = N_{obs}$ for τ^2 using PL and Storvik for a PoDLM. PMMH posterior mean and standard deviation included for comparison.

[1]Ground truth for state and parameter estimated using a PMCMC run (specifically PMMH, detailed in Chapter 13) for which the trace and ACF plots are available in Appendix A.2.

Algorithm 10.2 PL for non-linear DGLMs

initialisation

$\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0)$

$\boldsymbol{\Phi}_0 \sim p(\boldsymbol{\Phi})$

for $t \leftarrow 1$ to N_{obs}

for $i \leftarrow 1$ to N_p

calculate $\boldsymbol{\mu}_t^{(i)} = \mathrm{E}\left[\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, \Phi_{t-1}^{(i)}\right]$

calculate auxiliary weights

$$\lambda_t^{(i)} \propto p\left(y_t | \boldsymbol{\mu}_t^{(i)}, \boldsymbol{\Phi}_{t-1}^{(i)}\right) \tilde{w}_{t-1}$$

resample $\left\{s_{t-1}^{(i)\boldsymbol{\Phi}}\right\}_{i=1}^{N_p}$ according the auxiliary weights $\left\{\lambda_t^{(i)}\right\}_{i=1}^{N_p}$

for $i \leftarrow 1$ to N_p

sample $\boldsymbol{\theta}_t^{(i)} \sim p\left(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}^{\left(k^{(i)}\right)}, \boldsymbol{\Phi}_{t-1}^{\left(k^{(i)}\right)}\right)$

propagate parameter sufficient statistics according to

$$s_t^{\boldsymbol{\Phi}^{(i)}} = \mathcal{S}\left(s_{t-1}^{\boldsymbol{\Phi}^{\left(k^{(i)}\right)}}, \boldsymbol{\theta}_t^{(i)}, \boldsymbol{\theta}_{t-1}, y_t\right)$$

sample parameters according to

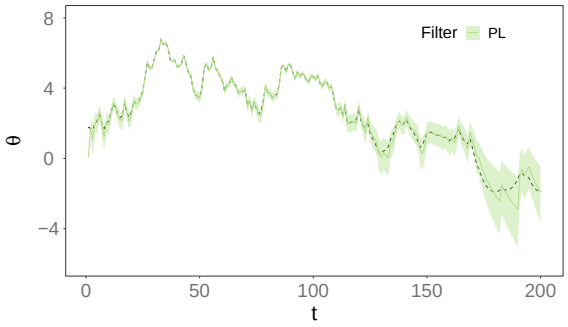
$$\boldsymbol{\Phi}^{(i)} \sim p\left(\boldsymbol{\Phi} | s_t^{\boldsymbol{\Phi}^{\left(k^{(i)}\right)}}\right)$$

calculate weights

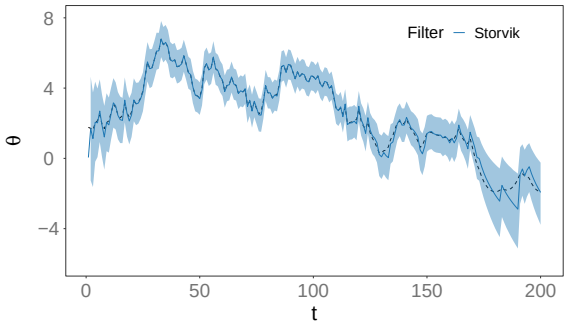
$$w_t^{(i)} \propto \frac{p\left(y_t | \boldsymbol{\theta}_t^{\left(k^{(i)}\right)}, \boldsymbol{\Phi}_t^{\left(k^{(i)}\right)}\right)}{p\left(y_t | \boldsymbol{\mu}_t^{(i)}, \boldsymbol{\Phi}_{t-1}^{(i)}\right)}$$

normalise weights

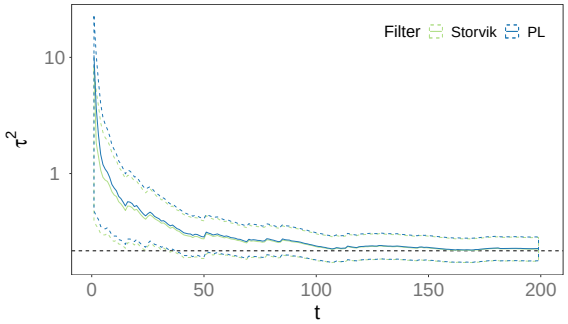
$$\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{i=1}^{N_p} w_t^{(i)}}$$



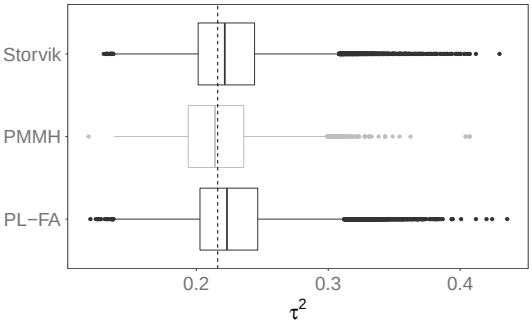
(a) $\theta_{0:t}$ posterior estimation using Storvik. Dashed line represents "true" state and shaded area the state posterior 90% equi-tailed credibility interval.



(b) $\theta_{0:t}$ posterior estimation using PL. Dashed line represents "true" state and shaded area the state posterior 90% equi-tailed credibility interval.



(c) τ^2 posterior estimation history. Horizontal dashed line represent the parameter's "true" value.



(d) $p\left(\tau^2|\mathcal{D}_t\right)$ posterior at $t = N_{obs}$. Vertical dashed line represents the parameter's "true" value.

Figure 10.2: State and parameter (τ^2) posterior estimation for a $\mathcal{M} = \{\mathcal{P}(1)\}$ PoDLM using $N_p = 10^5$ using the Storvik and PL filters.

Part IV

Offline State and Parameter Estimation

Chapter 11

Smoothing

11.1 Rauch–Tung–Striebel Smoother

So far we have considered the case of online filtering to determine the state posterior given the data up to time point t, that is $p(\boldsymbol{\theta}_t|\mathcal{D}_t)$. When dealing with Gaussian DLMs, this estimation can be performed using the Kalman Filter as described in Chapter 4. We might be interested, however, in estimating the state's posterior given the totality of the data $\mathcal{D}_T$, $p(\boldsymbol{\theta}_t|\mathcal{D}_T)$ for $t \leq T$, a method usually called *smoothing.*

In this chapter we will focus on the *Rauch-Tung-Strieble* (RTS) smoother, first introduced in Rauch *et al.* (1965). The RTS smoother is a fixed-interval smoother[1], meaning that the state's posterior estimates will be calculated given the whole of the data, $\mathcal{D}_T$. The RTS smoother provides the *optimal smoothing* solution for Gaussian DLMs and makes use of the state filtering estimations resulting from the Kalman Filter.

Considering, as in Chapter 4, that at time t the state estimate is given by the KF estimate in (4.13):

$$\boldsymbol{\theta}_t|\mathcal{D}_t \sim \mathcal{N}(\mathbf{m}_t, \mathsf{C}_t),$$

the RTS smoother will provide us with the state smoothed estimation given the totality of the data, that is:

$$\boldsymbol{\theta}_t|\mathcal{D}_T \sim \mathcal{N}\left(\tilde{\mathbf{m}}_t, \tilde{\mathsf{C}}_t\right), \qquad t \leq T, \tag{11.1}$$

where $\tilde{\mathbf{m}}_t$ and $\tilde{\mathsf{C}}_t$ represent the first and second moments respectively of the smoothed posterior at a given time $t \leq T$. To establish these moments, we can start by stating that

[1] Other smoother types include fixed-lag and fixed-point smoothing, not considered in this thesis.

the smoothed posterior is

$$p\left(\boldsymbol{\theta}_t|\mathcal{D}_T\right) = \int p\left(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right) d\boldsymbol{\theta}_{t+1}. \tag{11.2}$$

To determine the joint $p\left(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right)$ we can first determine the state's predictive density given the entirety of the data $\mathcal{D}_T$. Due to the Markovian nature of DGLMs described in (2.12) (that is $\boldsymbol{\theta}_t \perp\!\!\!\perp \left(\mathcal{D}_T \backslash \mathcal{D}_t\right) |\boldsymbol{\theta}_{t+1}$), it follows that

$$p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_T\right) = p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_t\right). \tag{11.3}$$

We can refactor (11.3) as

$$\begin{aligned} p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_T\right) &= p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_t\right) \\ &= \frac{p\left(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}|\mathcal{D}_t\right)}{p\left(\boldsymbol{\theta}_{t+1}|\mathcal{D}_t\right)} \\ &= \frac{p\left(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, \mathcal{D}_t\right) p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right)}{p\left(\boldsymbol{\theta}_{t+1}|\mathcal{D}_t\right)} \\ &= \frac{p\left(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t\right) p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right)}{p\left(\boldsymbol{\theta}_{t+1}|\mathcal{D}_t\right)}. \end{aligned} \tag{11.4}$$

Considering the smoothed posterior as in (11.1), we can consider the posterior at time $t+1$ as

$$\boldsymbol{\theta}_{t+1}|\mathcal{D}_T \sim \mathcal{N}\left(\tilde{\mathbf{m}}_{t+1}, \tilde{\mathsf{C}}_{t+1}\right), \qquad t+1 \leq T. \tag{11.5}$$

This allows us to rewrite the joint distribution $p\left(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right)$ as

$$\begin{aligned} p\left(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right) &= p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_T\right) p\left(\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right) \\ &= p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_t\right) p\left(\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right) \\ &= \frac{p\left(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t\right) p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right) p\left(\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right)}{p\left(\boldsymbol{\theta}_{t+1}|\mathcal{D}_t\right)}. \end{aligned}$$

Replacing this result in (11.2), we then have

$$\begin{aligned} p\left(\boldsymbol{\theta}_t|\mathcal{D}_T\right) &= \int \frac{p\left(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t\right) p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right) p\left(\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right)}{p\left(\boldsymbol{\theta}_{t+1}|\mathcal{D}_t\right)} d\boldsymbol{\theta}_{t+1} \\ &= p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right) \int \frac{p\left(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t\right) p\left(\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right)}{p\left(\boldsymbol{\theta}_{t+1}|\mathcal{D}_t\right)} d\boldsymbol{\theta}_{t+1}. \end{aligned} \tag{11.6}$$

For a Gaussian DLM, we seen in Chapter 4 that the filtering density is given by (4.13)

$$p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right) = \mathcal{N}\left(\boldsymbol{\theta}_t|\mathbf{m}_t, \mathsf{C}_t\right).$$

We can identify $p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, \mathcal{D}_t)$ as the predictive density (4.9) from Chapter (4) which is provided by

$$\begin{aligned} p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, \mathcal{D}_t) &= \mathcal{N}\left(\boldsymbol{\theta}_{t+1}|\underbrace{\mathsf{G}\mathbf{m}_t}_{\mathbf{a}_{t+1}}, \underbrace{\mathsf{G}\mathsf{C}_t\mathsf{G}^T + \mathsf{W}}_{\mathsf{R}_{t+1}}\right) \\ &= \mathcal{N}(\boldsymbol{\theta}_{t+1}|\mathbf{a}_{t+1}, \mathsf{R}_{t+1}). \end{aligned}$$

We can then express the joint posterior as

$$\begin{aligned} p(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}|\mathcal{D}_t) &= p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, \mathcal{D}_t)\, p(\boldsymbol{\theta}_t|\mathcal{D}_t) \\ &= \mathcal{N}(\boldsymbol{\theta}_{t+1}|\mathbf{a}_{t+1}, \mathsf{R}_{t+1})\, \mathcal{N}(\boldsymbol{\theta}_t|\mathbf{m}_t, \mathsf{C}_t). \end{aligned}$$

Using standard Normal theory, as described in (4.2), we can then write the joint density $p(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}|\mathcal{D}_t)$ as

$$\begin{aligned} p(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}|\mathcal{D}_t) &= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\theta}_t \\ \boldsymbol{\theta}_{t+1}\end{bmatrix}\middle|\begin{pmatrix}\mathrm{E}[\boldsymbol{\theta}_t|\mathcal{D}_t] \\ \mathrm{E}[\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t]\end{pmatrix}, \begin{pmatrix}\mathrm{Var}[\boldsymbol{\theta}_t|\mathcal{D}_t] & \mathrm{Cov}[\boldsymbol{\theta}_t, \boldsymbol{\theta}_t] \\ \mathrm{Cov}[\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}]^T & \mathrm{Var}[\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t]\end{pmatrix}\right) \\ &= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\theta}_t \\ \boldsymbol{\theta}_{t+1}\end{bmatrix}\middle|\begin{pmatrix}\mathbf{m}_t \\ \mathbf{a}_{t+1}\end{pmatrix}, \begin{pmatrix}\mathsf{C}_t & \mathsf{C}_t\mathsf{G}^T \\ \mathsf{G}\mathsf{C} & \mathsf{R}_{t+1}\end{pmatrix}\right). \end{aligned}$$

Since, as stated in (11.4), $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_T) = p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_t)$, by using the rules presented in (4.3) and (4.4) we can write (11.3) as

$$\begin{aligned} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_T) &= p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_t) \\ &= \mathcal{N}(\theta_t|\mathrm{E}[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_t], \mathrm{Var}[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_t]), \end{aligned}$$

where

$$\begin{aligned} \mathrm{E}[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathcal{D}_t] &= \mathrm{E}[\boldsymbol{\theta}_t|\mathcal{D}_t] + \mathrm{Cov}[\boldsymbol{\theta}_t, \boldsymbol{\theta}_t]\, \mathrm{Var}[\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t]^{-1} (\boldsymbol{\theta}_{t+1} - \mathrm{E}[\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t]) \\ &= \mathbf{m}_t + \underbrace{\mathsf{C}_t\mathsf{G}^T\mathsf{R}_{t+1}^{-1}}_{\mathsf{B}_t} (\boldsymbol{\theta}_{t+1} - \mathbf{a}_{t+1}) \qquad (11.7) \\ &= \mathbf{m}_t + \mathsf{B}_t (\boldsymbol{\theta}_{t+1} - \mathbf{a}_{t+1}), \qquad (11.8) \end{aligned}$$

with

$$\mathsf{B}_t = \mathsf{C}_t\mathsf{G}^T\mathsf{R}_{t+1}^{-1}, \qquad (11.9)$$

and

$$\begin{aligned}\text{Var}\left[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1},\mathcal{D}_t\right] &= \text{Var}\left[\boldsymbol{\theta}_t|\mathcal{D}_t\right] - \text{Cov}\left[\boldsymbol{\theta}_t,\boldsymbol{\theta}_t\right]\text{Var}\left[\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t\right]^{-1}\text{Cov}\left[\boldsymbol{\theta}_t,\boldsymbol{\theta}_{t+1}\right]^T \\ &= \mathsf{C}_t - \mathsf{B}_t\mathsf{R}_{t+1}\mathsf{B}_t^T. \end{aligned} \tag{11.10}$$

In order to write the joint posterior of $\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}$ given the entirety of the data, $p\left(\boldsymbol{\theta}_t,\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right)$ we can factor it as

$$p\left(\boldsymbol{\theta}_t,\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right) = p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1},\mathcal{D}_T\right)p\left(\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right).$$

We can now use the definition of $p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1},\mathcal{D}_T\right)$ in (11.8) and (11.10) and $p\left(\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right)$ from (11.5), such that

$$\begin{aligned}p\left(\boldsymbol{\theta}_t,\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right) &= \mathcal{N}\left(\mathbf{m}_t + \mathsf{B}_t\left(\boldsymbol{\theta}_{t+1} - \mathbf{a}_{t+1}\right), \mathsf{C}_t - \mathsf{B}_t\mathsf{R}_{t+1}\mathsf{B}_t^T\right) \\ &\quad \times\mathcal{N}\left(\tilde{\mathbf{m}}_{t+1}, \tilde{\mathsf{C}}_{t+1}\right).\end{aligned}$$

Using the definition of (11.8) and applying once more the rule in (4.3) we can then write the joint posterior $p\left(\boldsymbol{\theta}_t,\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right)$ as

$$\begin{aligned}p\left(\boldsymbol{\theta}_t,\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right) &= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\theta}_{t+1}\\ \boldsymbol{\theta}_t\end{bmatrix}\middle|\begin{pmatrix}\text{E}\left[\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right]\\ \text{E}\left[\boldsymbol{\theta}_t|\mathcal{D}_T\right]\end{pmatrix}, \begin{pmatrix}\text{Var}\left[\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right] & \text{Cov}\left[\boldsymbol{\theta}_t,\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right]\\ \text{Cov}\left[\boldsymbol{\theta}_t,\boldsymbol{\theta}_{t+1}|\mathcal{D}_T\right]^T & \text{Var}\left[\boldsymbol{\theta}_t|\mathcal{D}_T\right]\end{pmatrix}\right) \\ &= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\theta}_{t+1}\\ \boldsymbol{\theta}_t\end{bmatrix}\middle|\begin{pmatrix}\underbrace{\tilde{\mathbf{m}}_{t+1}}_{a}\\ \underbrace{\mathbf{m}_t + \mathsf{B}_t\left(\tilde{\mathbf{m}}_{t+1} - \mathbf{a}_{t+1}\right)}_{b}\end{pmatrix}, \mathcal{P}\right),\end{aligned}$$

with

$$\mathcal{P} = \begin{pmatrix}\underbrace{\tilde{\mathsf{C}}_{t+1}}_{A} & \underbrace{\tilde{\mathsf{C}}_{t+1}\mathsf{B}_t^T}_{C}\\ \underbrace{\mathsf{B}_t\tilde{\mathsf{C}}_{t+1}}_{C^T} & \underbrace{\mathsf{B}_t\tilde{\mathsf{C}}_{t+1}\mathsf{B}_t^T + \mathsf{C}_t - \mathsf{B}_t\mathsf{R}_{t+1}\mathsf{B}_t^T}_{B}\end{pmatrix}.$$

The marginal distribution $p\left(\boldsymbol{\theta}_t|\mathcal{D}_T\right)$ will then be

$$p\left(\boldsymbol{\theta}_t|\mathcal{D}_T\right) = \mathcal{N}\left(\tilde{\mathbf{m}}_t, \tilde{\mathsf{C}}_t\right),$$

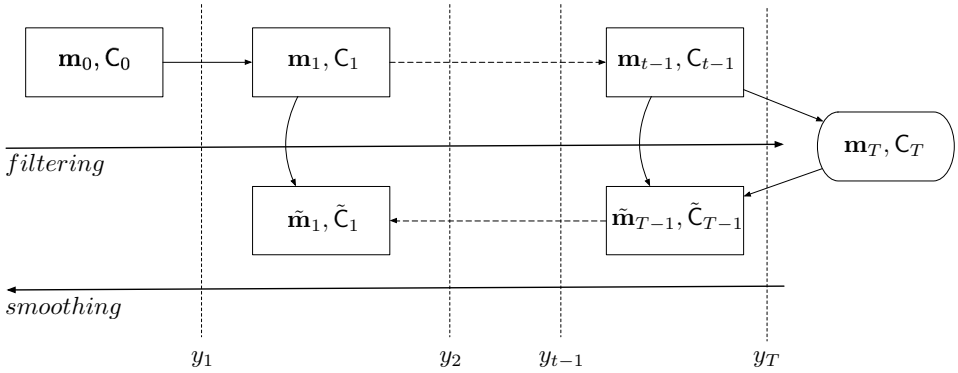


Figure 11.1: Illustration of the RTS smoother.

with

$$\begin{aligned} \tilde{\mathbf{m}}_t &= \mathbf{m}_t + \mathsf{B}_t \left(\tilde{\mathbf{m}}_{t+1} - \mathbf{a}_{t+1} \right) \\ \tilde{\mathsf{C}}_t &= \mathsf{C}_t + \mathsf{B}_t \left(\tilde{\mathsf{C}}_{t+1} - \mathsf{R}_{t+1} \right) \mathsf{B}_t^T. \end{aligned}$$

From this result, it is clear that the RTS smoother needs to make use, as mentioned previously, of the KF filter estimates (to provide $\mathbf{m}_t$ and C_t), proceeding in a backward recursion from the last time step $t = T$, as can be seen in Figure 11.1. As an initial condition for the smoother we define

$$\begin{aligned} \mathbf{m}_T &= \tilde{\mathbf{m}}_T \\ \mathsf{C}_T &= \tilde{\mathsf{C}}_T. \end{aligned}$$

The full algorithm for the RTS smoother is presented in Algorithm 11.1.

Example. RTS smoother for a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(75, 1)\}$ NDLM.

In this example we consider a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(75, 1)\}$ NDLM (period $p = 75$ for the seasonal component), with parameter set $\boldsymbol{\Phi} = \{W, V\} = \{\text{diag}(1.5, 0.1, 0.1), 4.3\}$. The initial state for the KF filter was $\theta_0 \sim \mathcal{N}\left(\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T, 10\mathbf{I}_3 \right)$. The full model specification

Algorithm 11.1 Rauch-Tung-Strieble (RTS) smoother

initialisation $(t = 0)$

Set $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0)$

For $t = 1 \rightarrow T$

Run the KF recursions according to Section 4.

Store $\{\mathsf{R}_t, \mathbf{m}_t, \mathsf{C}_t\}$

Set

$$\begin{aligned} \tilde{\mathbf{m}}_T &= \mathbf{m}_T \\ \tilde{\mathsf{C}}_T &= \mathsf{C}_T \end{aligned}$$

For $t = (T-1) \rightarrow 1$

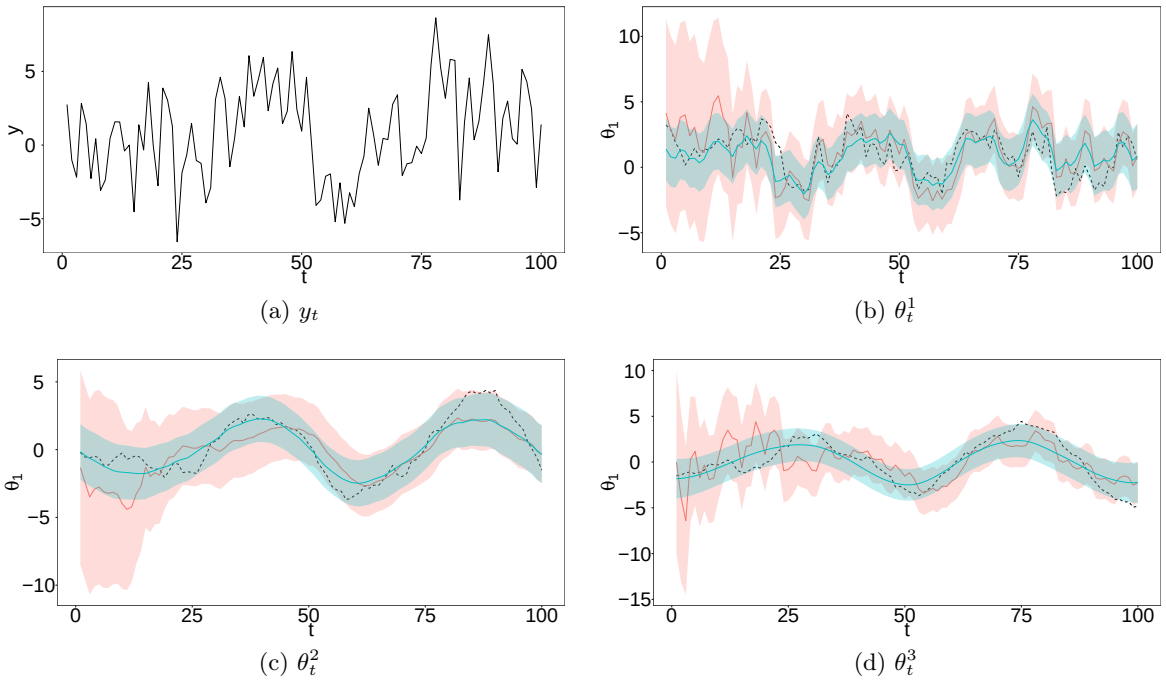
Calculate

$$\begin{aligned} \mathsf{B}_t &= \mathsf{C}_t \mathsf{G}^T \mathsf{R}_{t+1}^{-1} \\ \tilde{\mathbf{m}}_\mathbf{t} &= \mathbf{m}_t + \mathsf{B}_t \left(\tilde{\mathbf{m}}_{t+1} - \mathsf{G}\mathbf{m}_t\right) \\ \tilde{\mathsf{C}}_t &= \mathsf{C}_t + \mathsf{B}_t \left(\tilde{\mathsf{C}}_{t+1} - \mathsf{R}_{t+1}\right) \mathsf{B}_t^T \end{aligned}$$

is

$$\begin{aligned} y_t | \boldsymbol{\theta}_t, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\mathsf{F}^T \boldsymbol{\theta}_t, 4.3\right) \\ \boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\mathsf{G}\boldsymbol{\theta}_{t-1}, \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}\right) \\ \mathsf{F}^T &= \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \\ \mathsf{G} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\frac{2\pi}{75} & \sin\frac{2\pi}{75} \\ 0 & -\sin\frac{2\pi}{75} & \cos\frac{2\pi}{75} \end{bmatrix}. \end{aligned}$$

A realisation of $N_{obs} = 100$ was simulated and is presented in Figure 11.2a. An initial state estimation using the Kalman filter was performed and the resulting moment estimations, $\{\mathbf{m}_t, \mathsf{C}_t\}_{t=1}^T$, stored in order to proceed with the smoothed state estimation by the RTS smoother. The resulting estimates are presented in Figures 11.2b, 11.2c and 11.2d.



(a) y_t (b) θ_t^1

(c) θ_t^2 (d) θ_t^3

Figure 11.2: State components ($\{\theta_1, \theta_2, \theta_3\}$) estimated using the Kalman filter (*red*) and RTS smoother (*green*) and realisation's observations (*top left*) for a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(75, 1)\}$ Normal DLM. Solid colour lines represent the filtering/smoothing density mean and shaded areas the 90% CI. Dashed line represents the realisation's state mean.

Chapter 12

Expectation-Maximisation

To estimate the parameter set $\boldsymbol{\Phi} = \{\mathsf{W}, V\}$ in a NDLM, as specified in Section 2.3.1, we can apply the *Expectation-Maximisation*(EM) method as described in Shumway & Stoffer (1982). The EM algorithm is an iterative algorithm used to estimate unknown parameters and can be considered as specific class of maximum likelihood estimation methods (MLE).

If we consider the augmented data as

$$\mathcal{Y}_T = \{\mathcal{D}_T, \Theta_T\}, \tag{12.1}$$

with, as previously, $\mathcal{D}_T = \{y_1, y_2, \ldots, y_T\}$ and $\Theta_T = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T\}$, the EM aims at (starting from an initial parameter set $\boldsymbol{\Phi}_0 = \{\mathsf{W}_0, V_0\}$) iteratively maximising the conditional expectation of the log-likelihood of (12.1) conditioned on the total data, $\mathcal{D}_T$, and the previous parameter set, that is

$$\mathrm{E}_{\Theta_T}\left[\ell\left(\mathcal{Y}_T|\boldsymbol{\Phi}\right)|\mathcal{D}_T, \boldsymbol{\Phi}^{(j-1)}\right] \qquad j = 1, 2, \ldots \tag{12.2}$$

The EM algorithm consists of two steps, namely the Expectation Step (E-step) and the Maximisation Step (M-step). The E-step consists of calculating the expectation in (12.2). In the second step, we maximise the result obtained in the E-step conditioned on $\boldsymbol{\Phi}$ by calculating

$$\boldsymbol{\Phi}^{(j)} \leftarrow \arg\max_{\Phi} \mathrm{E}_{\Theta_T}\left[\ell\left(\mathcal{Y}_T|\boldsymbol{\Phi}\right)|\mathcal{D}_T, \boldsymbol{\Phi}^{(j-1)}\right],$$

These steps are iteratively computed until some convergence criteria is reached, in essence interpreting the parameter estimation as a model optimisation. A typical criteria will depend on the actual relative change of the parameters when compared to a minimum threshold ϵ. For instance, if we consider the parameter set $\boldsymbol{\Phi} = \{\Phi_1, \ldots, \Phi_n\}$, at a certain

iteration step j, the relative change will be

$$\delta\boldsymbol{\Phi} = \max_i \left\{ \frac{\left| \boldsymbol{\Phi}_i^{(j)} - \boldsymbol{\Phi}_i^{(j-1)} \right|}{\boldsymbol{\Phi}_i^{(j)}} \right\}_{i=1}^{n},$$

implying a stopping criterion of

$$\delta\boldsymbol{\Phi} \leq \epsilon.$$

Each iteration of the EM algorithm is guaranteed to increase the marginal log-likelihood. It will therefore converge to a local mode. When applied to the NDLM case, the EM method guarantees convergence according to Wu (1983).

To define the EM algorithm for NDLM, we must first be able to calculate the conditional expectation in (12.2), which for a Normal DLM we can do analytically. From Section 2.3.1 we know that the general Normal DLM can be written in the form

$$\begin{aligned} y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\mathsf{F}^T\boldsymbol{\theta}_t, V\right) \\ \boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}\right). \end{aligned}$$

Using the definition for the Normal PDF for the multivariate case in (4.1), we can then reformulate the above NDLM as

$$\begin{aligned} p\left(y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi}\right) &= \frac{1}{\sqrt{(2\pi)^k \left|\mathsf{V}\right|}} \exp\left\{ -\frac{1}{2}\left(y_t - \mathsf{F}^T\boldsymbol{\theta}_t\right)^T \mathsf{V}^{-1}\left(y_t - \mathsf{F}^T\boldsymbol{\theta}_t\right) \right\} \\ p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi}\right) &= \frac{1}{\sqrt{(2\pi)^k \left|\mathsf{W}\right|}} \exp\left\{ -\frac{1}{2}\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right)^T \mathsf{W}^{-1}\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right) \right\} \end{aligned}$$

We will use here the multivariate version for the observation equation. This is will not alter the final result since this is simply a general formulation of the univariate case.

The likelihood of the augmented data can be factorised as

$$p\left(\mathcal{Y}_T\right) = p\left(\mathcal{D}_T, \Theta_T\right) = p\left(\mathcal{D}_T|\Theta_T\right) p\left(\Theta_T\right).$$

We know that

$$p\left(\Theta_T\right) = p\left(\boldsymbol{\theta}_0\right) \prod_{t=1}^{T} p\left(\boldsymbol{\theta}_t|\Theta_{t-1}\right) \tag{12.3}$$

$$p\left(\mathcal{D}_T|\Theta_T\right) = \prod_{t=1}^{T} p\left(y_t|\Theta_T\right). \tag{12.4}$$

However, due to the Markovian nature of DGLMs, as specified in Section 2.1, we can rewrite (12.3) and (12.4) as

$$\begin{aligned} p\left(\Theta_T\right) &= p\left(\boldsymbol{\theta}_0\right) \prod_{t=1}^{T} p\left(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}\right) \\ p\left(\mathcal{D}_T | \Theta_T\right) &= \prod_{t=1}^{T} p\left(y_t | \boldsymbol{\theta}_t\right). \end{aligned}$$

Consequently, the likelihood of the augmented data in (12.1) can be written as

$$\mathcal{L}\left(\mathcal{Y}_T | \boldsymbol{\Phi}\right) = p\left(\boldsymbol{\theta}_0\right) \prod_{t=1}^{T} p\left(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}\right) p\left(y_t | \boldsymbol{\theta}_t\right).$$

Since the we consider an initial state prior also in the Normal form, that is

$$\boldsymbol{\theta}_0 \sim \mathcal{N}\left(\mathbf{m}_0, \mathsf{C}_0\right)$$

we also consider $p\left(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_0\right) = p\left(\boldsymbol{\theta}_1 | \mathbf{m}_0, \mathsf{C}_0\right)$ to have the Normal form, that is

$$p\left(\boldsymbol{\theta}_1\right) = \frac{1}{\sqrt{\left(2\pi\right)^k \left|\mathsf{C}_0\right|}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}_\mathbf{1} - \mathbf{m}_0\right)^T \mathsf{C}_0^{-1}\left(\boldsymbol{\theta}_t - \mathbf{m}_0\right)\right\}.$$

We can further simplify the likelihood of the augmented data by working with the log-likelihood, that is calculating

$$\begin{aligned} \ell\left(\mathcal{Y}_T | \boldsymbol{\Phi}\right) &= \log\left\{\mathcal{L}\left(\mathcal{Y}_T | \boldsymbol{\Phi}\right)\right\} \\ &= -\frac{1}{2}\left|\mathsf{C}_0\right| - \frac{1}{2}\left(\boldsymbol{\theta}_0 - \mathbf{m}_0\right)^T \mathsf{C}_0^{-1}\left(\boldsymbol{\theta}_0 - \mathbf{m}_0\right) \\ &\quad -\frac{T}{2}\log\left|\mathsf{W}\right| - \frac{1}{2}\sum_{t=1}^{T}\left\{\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right)^T \mathsf{W}^{-1}\left(\boldsymbol{\theta}_t - \mathsf{G}\boldsymbol{\theta}_{t-1}\right)\right\} \\ &\quad -\frac{T}{2}\log\left|\mathsf{V}\right| - \frac{1}{2}\sum_{t=1}^{T}\left\{\left(y_t - \mathsf{F}^T\boldsymbol{\theta}_t\right)^T \mathsf{V}^{-1}\left(y_t - \mathsf{F}^T\boldsymbol{\theta}_t\right)\right\} \end{aligned}$$

If we express the conditional expectation in (12.2) as function of $\widehat{\boldsymbol{\Phi}}^{(j)}$ (the estimated parameter set at iteration j), we have

$$Q\left(\boldsymbol{\Phi}, \widehat{\boldsymbol{\Phi}}^{(j)}\right) = \mathrm{E}\left[\ell\left(\mathcal{Y}_T | \boldsymbol{\Phi}\right) | \mathcal{D}_T, \widehat{\boldsymbol{\Phi}}^{(j)}\right]. \tag{12.5}$$

According to Shumway & Stoffer (1982), by defining the conditional mean and covariances

$$\begin{aligned}\tilde{\mathbf{m}}_t &= \mathrm{E}\left[\boldsymbol{\theta}_t|\mathcal{D}_T\right] \\ \tilde{\mathsf{C}}_t &= \mathrm{Cov}\left[\boldsymbol{\theta}_t|\mathcal{D}_T\right] \\ \tilde{\mathsf{C}}_{t|t-1} &= \mathrm{Cov}\left[\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}|\mathcal{D}_T\right],\end{aligned}$$

we can further simplify (and using the notation in Shumway & Stoffer (1982)) the conditional expectation on 12.2 as

$$\begin{aligned}Q\left(\boldsymbol{\Phi}, \widehat{\boldsymbol{\Phi}}^{(j)}\right) = & -\frac{1}{2}\left|\mathsf{C}_0\right| - \frac{1}{2}\mathrm{tr}\left\{\mathsf{C}_0^{-1}\left(\tilde{\mathsf{C}}_0 + \left(\tilde{\mathbf{m}}_0 - \mathbf{m}_0\right)\left(\tilde{\mathbf{m}}_0 - \mathbf{m}_0\right)^T\right)\right\} \\ & -\frac{T}{2}\log\left|\mathsf{W}\right| - \frac{1}{2}\mathrm{tr}\left\{\mathsf{W}^{-1}\left(\mathcal{C} - \mathcal{B}\mathsf{G}^T - \mathsf{G}\mathcal{B}^T + \mathsf{G}\mathcal{A}\mathsf{G}^T\right)\right\} \\ & -\frac{T}{2}\log\left|\mathsf{V}\right| \\ & -\frac{1}{2}\mathrm{tr}\left\{\mathsf{V}^{-1}\sum_{t=1}^{T}\left[\left(y_t - \mathsf{F}^T\tilde{\mathbf{m}}_t\right)\left(y_t - \mathsf{F}^T\tilde{\mathbf{m}}_t\right)^T + \mathsf{F}^T\tilde{\mathsf{C}}_t\mathsf{F}\right]\right\} \quad (12.6)\end{aligned}$$

where

$$\begin{aligned}\mathcal{A} &= \sum_{t=1}^{T}\left(\tilde{\mathsf{C}}_{t-1} + \tilde{\mathbf{m}}_{t-1}\tilde{\mathbf{m}}_{t-1}^T\right) \\ \mathcal{B} &= \sum_{t=1}^{T}\left(\tilde{\mathsf{C}}_{t|t-1} + \tilde{\mathbf{m}}_t\tilde{\mathbf{m}}_{t-1}^T\right) \\ \mathcal{C} &= \sum_{t=1}^{T}\left(\tilde{\mathsf{C}}_t + \tilde{\mathbf{m}}_t\tilde{\mathbf{m}}_t^T\right)\end{aligned}$$

and $\mathrm{tr}(\cdot)$ is the trace function. As noted in Shumway & Stoffer (1982), to calculate the analytical form of (12.6), we made extensive use of the identities derived during the RTS smoother analysis in Chapter 11 on page 137, since we are calculating the expectations in (12.5) of, for instance, the states $\boldsymbol{\theta}_t$, not conditioned on the data up until time t, but rather conditioned on the entirety of the data, that is $\boldsymbol{\theta}_t|\mathcal{D}_T$. This is will be, by definition, a smoothing problem and the calculation of (12.5) will rely namely on smoothing quantities

$$\begin{aligned}\tilde{\mathbf{m}}_t &= \mathrm{E}\left[\boldsymbol{\theta}_t|\mathcal{D}_T\right] \\ \tilde{\mathsf{C}}_t &= \mathrm{E}\left[\boldsymbol{\theta}_t\boldsymbol{\theta}_t^T|\mathcal{D}_T\right] \\ \tilde{\mathsf{C}}_{t-1} &= \mathrm{E}\left[\boldsymbol{\theta}_{t-1}\boldsymbol{\theta}_{t-1}^T|\mathcal{D}_T\right] \\ \tilde{\mathsf{C}}_{t|t-1} &= \mathrm{E}\left[\boldsymbol{\theta}_t\boldsymbol{\theta}_{t-1}|\mathcal{D}_T\right].\end{aligned}$$

It is clear from (12.6) that for the E-step, not only we will need the values provided

by RTS smoother for the expectation calculation, but also, implicitly the KF estimates so the smoother can be evaluated.

The maximisation step (M-step) will consist in maximising the conditional expectation in (12.6). Since the logarithm is a monotone function, that is

$$\arg\max\left\{\log\left[f\left(\cdot\right)\right]\right\} = \arg\max\left\{f\left(\cdot\right)\right\},$$

we can calculate the first order partial derivative of (12.6) with regard to the parameters and set it to zero. Solving this in function of parameters will provide the values that maximise (12.6). As such, we calculate

$$\begin{aligned}\frac{\partial Q\left(\boldsymbol{\Phi}, \widehat{\boldsymbol{\Phi}}^{(j)}\right)}{\partial \mathsf{W}} &= 0\\ \frac{\partial Q\left(\boldsymbol{\Phi}, \widehat{\boldsymbol{\Phi}}^{(j)}\right)}{\partial \mathsf{V}} &= 0.\end{aligned}$$

For the W we can simplify the calculation by calculating the derivative with regard to W^{-1} such that

$$\begin{aligned}\frac{\partial Q\left(\boldsymbol{\Phi}, \widehat{\boldsymbol{\Phi}}^{(j)}\right)}{\partial \mathsf{W}^{-1}} &= 0\\ \frac{T}{2}\mathsf{W} - \frac{1}{2}\left(\sum_{t=0}^{T}\left(\tilde{\mathbf{m}}_t - \mathsf{G}_t\tilde{\mathbf{m}}_{t-1}\right)\left(\tilde{\mathbf{m}}_t - \mathsf{G}_t\tilde{\mathbf{m}}_{t-1}\right)^T + \mathsf{L}_t\right)^T &= 0,\end{aligned}$$

where

$$\mathsf{L}_t = \tilde{\mathsf{C}}_t + \mathsf{G}\tilde{\mathsf{C}}_{t-1}\mathsf{G}^T - \tilde{\mathsf{C}}_t\mathsf{B}_{t-1}^T\mathsf{G}^T - \mathsf{G}\mathsf{B}_{t-1}\tilde{\mathsf{C}}_t^T,$$

and B_t as defined in (11.9). Solving in order of W results in the new parameter estimate at step j

$$\mathsf{W}^{(j+1)} = \frac{1}{T}\left(\sum_{t=0}^{T}\left(\tilde{\mathbf{m}}_t - \mathsf{G}_t\tilde{\mathbf{m}}_{t-1}\right)\left(\tilde{\mathbf{m}}_t - \mathsf{G}_t\tilde{\mathbf{m}}_{t-1}\right)^T + \mathsf{L}_t\right),$$

which maximises (12.6) with regard to W. With respect to V, we calculate

$$\begin{aligned}\frac{\partial Q\left(\boldsymbol{\Phi}, \widehat{\boldsymbol{\Phi}}^{(j)}\right)}{\partial V^{-1}} &= 0\\ \frac{T+1}{2}V - \frac{1}{2}\left(\sum_{t=0}^{T}\left\{\left(y_t - \mathsf{F}^T\tilde{\mathbf{m}}_t\right)\left(y_t - \mathsf{F}^T\tilde{\mathbf{m}}_t\right)^T + \mathsf{F}^T\tilde{\mathsf{C}}_t\mathsf{F}\right\}\right)^T &= 0,\end{aligned}$$

therefore the value which maximises (12.6) with regard to V is

$$V^{(j+1)} = \frac{1}{T+1}\left(\sum_{t=0}^{T}\left\{\left(y_t - \mathsf{F}^T\tilde{\mathbf{m}}_t\right)\left(y_t - \mathsf{F}^T\tilde{\mathbf{m}}_t\right)^T + \mathsf{F}^T\tilde{\mathsf{C}}_t\mathsf{F}\right\}\right).$$

Some limitations of the general EM method include the fact that it stabilises when converging to a local maximum of the likelihood (which might not necessarily be the global maximum) and the fact that it only provides a point estimate of the the parameter set (although, according to Shumway & Stoffer (1982), perturbing the likelihood near the maximum may allow for the calculation of standard errors) which maximises the likelihood.

The full EM algorithm is defined in Algorithm 12.1.

Example. EM for a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(200, 1)\}$ Normal DLM

For a simulation of the EM parameter estimation, we have chosen a Normal DLM with a locally constant component composed with seasonality. The locally constant component had a state variance $W = 1.5$ and the seasonality consisted of $h = 1$ harmonics with a period $p = 200$ and a variance of $\mathsf{W} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. The observation variance was $V = 1.11$. We can see the data for a realisation for $N_{obs} = 2000$ of this model in Figure 12.1a. This corresponds to a model in the form

$$\begin{aligned} y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\mathsf{F}^T\boldsymbol{\theta}_t, 1.1\right) \\ \boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\mathsf{G}\boldsymbol{\theta}_{t-1}, \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right). \end{aligned}$$

The initial parameters for the estimation were set at

$$V_0 = 10, \qquad \mathsf{W}_0 = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{bmatrix}.$$

The EM algorithm converged after 902 iterations using $\epsilon = 0.001$ and the final values were $\tilde{V} = 1.210648$ and $\text{diag}\left(\tilde{\mathsf{W}}\right) = \begin{bmatrix} 1.7400 & 0.9765 & 1.3361 \end{bmatrix}$, with a running time of 9.102 seconds. The estimation values for each iteration can be viewed in Figure 12.1.

Algorithm 12.1 Expectation-Maximisation (for NDLM)

initialisation

set a value for $\boldsymbol{\Phi}^{(0)} = \left\{\mathsf{W}^{(0)}, V^{(0)}\right\}$

set a convergence threshold ϵ

set $j = 0$

while $\delta\boldsymbol{\Phi} > \epsilon$

(E-step)

calculate $\{\mathbf{m}_t, \mathsf{C}_t\}_{t=1}^{T}$ using the Kalman Filter conditioned on $\boldsymbol{\Phi}^{(j)}$

calculate $\left\{\tilde{\mathbf{m}}_t, \tilde{\mathsf{C}}_t\right\}_{t=1}^{T}$ using the RTS smoother conditioned on $\boldsymbol{\Phi}^{(j)}$

(M-step)

calculate the new parameter $\mathsf{W}^{(j+1)}$

$$\mathsf{W}^{(j+1)} = \frac{1}{T}\left(\sum_{t=0}^{T}\left(\tilde{\mathbf{m}}_t - \mathsf{G}_t\tilde{\mathbf{m}}_{t-1}\right)\left(\tilde{\mathbf{m}}_t - \mathsf{G}_t\tilde{\mathbf{m}}_{t-1}\right)^T + \mathsf{L}_t\right)$$

calculate the new parameter $V^{(j+1)}$

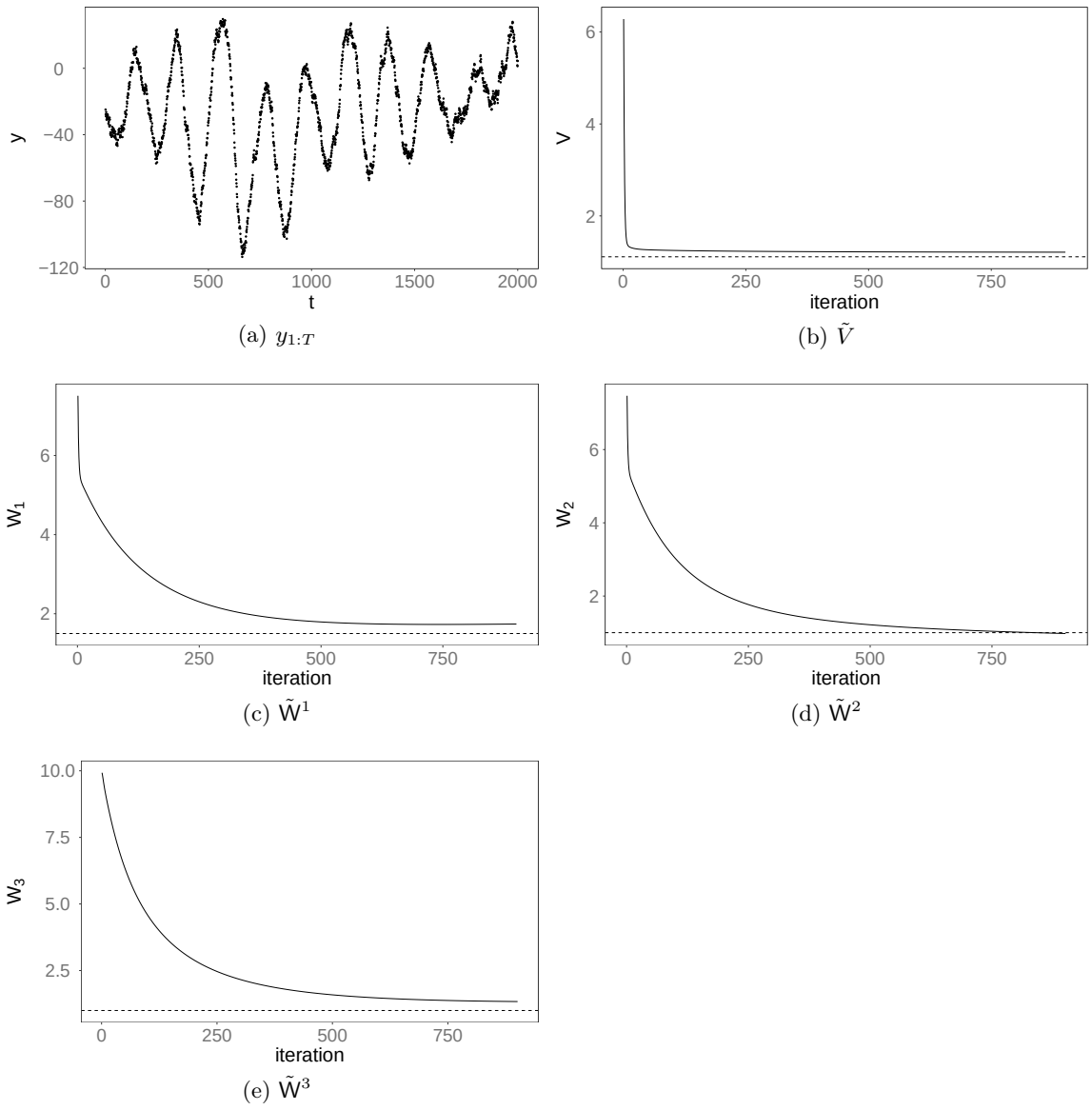
$$V^{(j+1)} = \frac{1}{T+1}\left(\sum_{t=0}^{T}\left\{\left(y_t - \mathsf{F}^T\tilde{\mathbf{m}}_t\right)\left(y_t - \mathsf{F}^T\tilde{\mathbf{m}}_t\right)^T + \mathsf{F}^T\tilde{\mathsf{C}}_t\mathsf{F}\right\}\right)$$

set $\boldsymbol{\Phi}^{(j+1)} = \left\{\mathsf{W}^{(j+1)}, V^{(j+1)}\right\}$

set $j \leftarrow j + 1$

calculate relative change

$$\delta\boldsymbol{\Phi} = \max_i\left\{\frac{\left|\boldsymbol{\Phi}_i^{(j)} - \boldsymbol{\Phi}_i^{(j-1)}\right|}{\boldsymbol{\Phi}_i^{(j)}}\right\}_{i=1}^{n}$$



(a) $y_{1:T}$ (b) $\tilde{V}$

(c) $\tilde{\mathsf{W}}^1$ (d) $\tilde{\mathsf{W}}^2$

(e) $\tilde{\mathsf{W}}^3$

Figure 12.1: Observations (*top left*) and EM estimation history (for $n = 902$ iterations) of $\boldsymbol{\Phi} = \{\mathsf{W}, V\}$ for a realisation of a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(200, 1)\}$ Normal DLM with $N_{obs} = 2000$. Horizontal dashed line represents true parameter value.

Chapter 13

Particle Markov Chain Monte Carlo

An additional class of methods to perform offline state and parameter estimation is Particle Markov Chain Monte Carlo (PMCMC), introduced in Andrieu *et al.* (2010). PMCMC methods represent a valuable tool to perform inference in state-space models with proven theoretical results and have been chosen to be the "gold standard" for subsequent comparisons in this thesis.

If we consider, in a state-space model, the joint distribution of parameter, states and observations, $p(\boldsymbol{\Phi}, \boldsymbol{\theta}, \mathcal{D}_T)$, we can factor it as

$$p(\boldsymbol{\Phi}, \boldsymbol{\theta}, \mathcal{D}_T) = p(\boldsymbol{\Phi})\, p(\boldsymbol{\theta}|\boldsymbol{\Phi})\, p(\mathcal{D}_T|\boldsymbol{\theta}, \boldsymbol{\Phi}). \tag{13.1}$$

A standard approach to estimate the parameter posterior $p(\boldsymbol{\Phi}|\mathcal{D}_t)$ would be to use Markov Chain Monte Carlo (MCMC) methods. The states can in principle be marginalised in order to obtain the observation's marginal likelihood, such that

$$p(\mathcal{D}_T|\boldsymbol{\Phi}) = \int p(\boldsymbol{\theta}|\boldsymbol{\Phi})\, p(\mathcal{D}_T|\boldsymbol{\theta}, \boldsymbol{\Phi})\, d\boldsymbol{\theta}, \tag{13.2}$$

and the (ideal) Metropolis-Hasting method could be used to estimate the parameters. The Metropolis-Hastings (MH) algorithm is a fundamental (and possibly the most popular MCMC method; Sherlock *et al.* (2015)), where assuming we want to estimate a target distribution

$$\pi(\boldsymbol{\Phi}) = p(\boldsymbol{\Phi}|\mathcal{D}_T),$$

the MH algorithm will consist (in general terms) of sampling a proposed parameter $\boldsymbol{\Phi}^\star$ from a proposal distribution, conditioned on a previously accepted parameter, $\boldsymbol{\Phi}'$, such that

$$\boldsymbol{\Phi}^\star \sim q\left(\boldsymbol{\Phi}^\star|\boldsymbol{\Phi}'\right).$$

After calculating an acceptance probability A as

$$\begin{aligned} A &= 1 \wedge \frac{\pi\left(\boldsymbol{\Phi}^{\star}\right)}{\pi\left(\boldsymbol{\Phi}'\right)} \frac{q\left(\boldsymbol{\Phi}'|\boldsymbol{\Phi}^{\star}\right)}{q\left(\boldsymbol{\Phi}^{\star}|\boldsymbol{\Phi}'\right)} \\ &= 1 \wedge \frac{p\left(\boldsymbol{\Phi}^{\star}\right)}{p\left(\boldsymbol{\Phi}'\right)} \frac{p\left(\mathcal{D}_T|\boldsymbol{\Phi}^{\star}\right)}{p\left(\mathcal{D}_T|\boldsymbol{\Phi}^{\star}\right)} \frac{q\left(\boldsymbol{\Phi}'|\boldsymbol{\Phi}^{\star}\right)}{q\left(\boldsymbol{\Phi}^{\star}|\boldsymbol{\Phi}'\right)}, \end{aligned} \tag{13.3}$$

we either accept the new parameter $\boldsymbol{\Phi}^{\star}$ with probability A or keep the previous value $\boldsymbol{\Phi}'$. Firstly, we can see from (13.3) that we must be able to calculate the marginal likelihoods $p\left(\mathcal{D}_T|\cdot\right)$. Secondly, we can also see that by using a symmetric proposal distribution, that is

$$q\left(\boldsymbol{\Phi}'|\boldsymbol{\Phi}^{\star}\right) = q\left(\boldsymbol{\Phi}^{\star}|\boldsymbol{\Phi}'\right),$$

we can simplify the calculation of the acceptance ratio into

$$A = 1 \wedge \frac{p\left(\boldsymbol{\Phi}^{\star}\right)}{p\left(\boldsymbol{\Phi}'\right)} \frac{p\left(\mathcal{D}_T|\boldsymbol{\Phi}^{\star}\right)}{p\left(\mathcal{D}_T|\boldsymbol{\Phi}^{\star}\right)}.$$

A typical[1] proposal distribution is a multivariate normal random walk centred on the previous parameter values $\boldsymbol{\Phi}'$ with a defined variance $\boldsymbol{\Sigma}$, such that

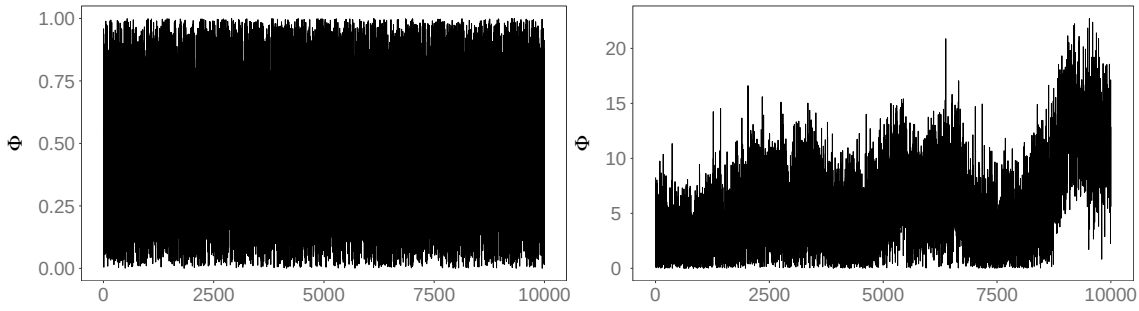
$$\begin{aligned} \boldsymbol{\Phi}^{\star} &\sim q\left(\boldsymbol{\Phi}^{\star}|\boldsymbol{\Phi}'\right) \\ &\sim \mathcal{N}\left(\boldsymbol{\Phi}^{\star}; \boldsymbol{\Phi}', \boldsymbol{\Sigma}\right). \end{aligned}$$

Such a distribution is symmetrical and allows for a simplified acceptance ratio calculation. In possession of the likelihood (13.2), new parameters $\boldsymbol{\Phi}^{\star}$ could be estimated using a (general) Metropolis-Hastings algorithm such as presented in Algorithm 13.1 on page 155. As we can see from (13.3), the MH method allows parameter proposals to be accepted when they cause the target (the posterior in this case) to be higher or if we have the same probability, if the likelihood change is small enough, enabling the sampler to visit regions of higher likelihood more frequently. To quantify the efficiency of the MH method, two heuristic measures are useful, namely the *auto-correlation function* (ACF) and the *effective sample size*. According to Liu (2002) the lag-k ACF for a stationary stochastic process $\{\mathcal{M}_t\}_{t=1}^{T}$ is defined by

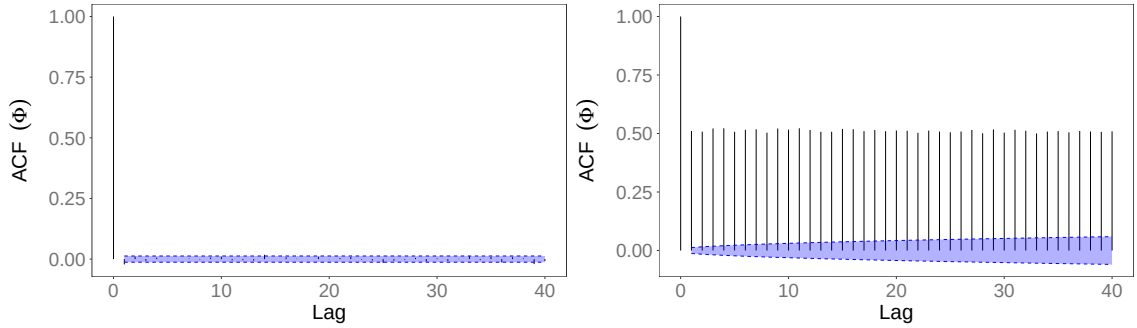
$$\rho_k = \operatorname{corr}\left(\mathcal{M}_{(1)}, \mathcal{M}_{(k+1)}\right). \tag{13.4}$$

Lower auto-correlations are of course desirable, since this will indicate a lower amount

[1] According to conventional MCMC theory (Roberts & Rosenthal (2001)), the optimal proposal step for a $\mathcal{N}\left(\mathbf{0}, \mathbf{I}_p\right)$ target can be set to $\boldsymbol{\Sigma}_p = \left(\lambda^2/p\right)\mathbf{I}_p$ with $\lambda = 2.38$.



(a) Trace of MH samples



(b) k-lag ACF of MH samples

Figure 13.1: MH sampler trace and k-lag ACF for simulated $\boldsymbol{\Phi}$ parameters representing low correlation (*left*) and high correlation (*right*)

of correlated (hence dependent) samples. A measure closely related to the ACF, is the (Monte Carlo) effective sample size (MCESS, Brooks *et al.* (2011)), which will provide an heuristic estimate for the number of independent samples produced. The MCESS can be defined as

$$MCESS = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k},$$

where n is the total number of samples and ρ_k is defined in (13.4). It is clear that as the auto-correlation increases, MCESS will be lower, whereas in the limit if all the samples are independent (implying $\rho_k = 0$ for $k = 1, \ldots, \infty$), then the MCESS would be n, the totality of the samples. An illustration of the quantities applied to simulated $\boldsymbol{\Phi}$ parameters with respectively low and high correlation are presented in Figure 13.1. Two additional approaches can be used when performing posterior estimation, namely *burn-in* and *thinning*. Burn-in is performed by removing the initial portion of the samples, that is, the initial values of the chain before it converges to a stationary region. This approach aims at minimising the impact of those samples in the final posterior estimation. Thinning is performed by only keeping every n^{th} sample, with $n > 1$, for the posterior estimation. The main goal is reducing the storage costs (however, if the parameters are highly correlated, thinning can potentially help with producing a less dependent final set of samples).

Algorithm 13.1 Metropolis-Hastings

initialise $n = 1$

$\boldsymbol{\Phi}^{(1)} \sim p(\boldsymbol{\Phi})$

for $n \geq 1$

propose $\boldsymbol{\Phi}^{\star} \sim q\left(\cdot|\boldsymbol{\Phi}^{(n-1)}\right)$

calculate

$$A = 1 \wedge \frac{q\left(\boldsymbol{\Phi}^{(n-1)}|\boldsymbol{\Phi}^{\star}\right) p(\boldsymbol{\Phi}^{\star}) p(\mathcal{D}_T|\boldsymbol{\Phi}^{\star})}{q\left(\boldsymbol{\Phi}^{\star}|\boldsymbol{\Phi}^{(n-1)}\right) p\left(\boldsymbol{\Phi}^{(n-1)}\right) p\left(\mathcal{D}_T|\boldsymbol{\Phi}^{(n-1)}\right)}$$

accept $\boldsymbol{\Phi}^{(n)} = \boldsymbol{\Phi}^{\star}$ with probability A or reject $\boldsymbol{\Phi}^{(n)} = \boldsymbol{\Phi}^{(n-1)}$.

In the context of this thesis, especially with non-linear DGLMs, $p(\mathcal{D}_T|\boldsymbol{\Phi})$ is not available. However, as we have seen in Section 6.1, SMC methods (such as SIR) can be used as an unbiased estimator of $p(\mathcal{D}_T|\boldsymbol{\Phi})$ by calculating

$$\begin{aligned} p(\mathcal{D}_T|\boldsymbol{\Phi}) &\approx \prod_{t=1}^{T} \frac{1}{N_p} \sum_{i=1}^{N_p} \tilde{w}_t^{(i)} \\ &\approx \widehat{p}(\mathcal{D}_T|\boldsymbol{\Phi}). \end{aligned}$$

PMCMC methods build on the principle of using the SMC approximation $\widehat{p}(\mathcal{D}_T|\boldsymbol{\Phi})$ to approximate $p(\mathcal{D}_T|\boldsymbol{\Phi})$. To better illustrate PMCMC we start with describing an implementation known as the Particle Independent Metropolis-Hastings (PIMH) following the algorithm presented in 13.2 on page 157. This method, presented in Andrieu *et al.* (2010) aims at estimating $p(\boldsymbol{\theta}_{0:T}|\mathcal{D}_T, \boldsymbol{\Phi})$, that is, we assume the parameter set $\boldsymbol{\Phi}$ is known. This method, as noted in Andrieu *et al.* (2010), although not a serious competitor to SMC state estimation, still provides a valuable insight on the workings of PMCMC methods.

We start by introducing a notation for a generic particle filter where

$$\left\{\boldsymbol{\theta}_{0:T}^{(i)}, a_{1:T}^{(i)}, w_{0:T}^{(i)}\right\}_{i=1}^{N_p} \sim \mathcal{PF}(\mathcal{D}_T, \boldsymbol{\Phi}).$$

As we have seen from Section 6.1, a general particle filter will at time t provide an approximation to the marginal posterior $p(\boldsymbol{\theta}_t|\mathcal{D}_t)$ with a discrete set of particles and weights $\left\{\boldsymbol{\theta}_t^{(i)}, w_t^{(i)}\right\}_{i=1}^{N_p}$. At time $t+1$, the approximation will be $\left\{\boldsymbol{\theta}_{t+1}^{(i)}, w_{t+1}^{(i)}\right\}_{i=1}^{N_p}$. The *ancestors* of the particle set at time $t+1$ will be direct mapping from the each particle $\boldsymbol{\theta}_t^{(i)}$ to the

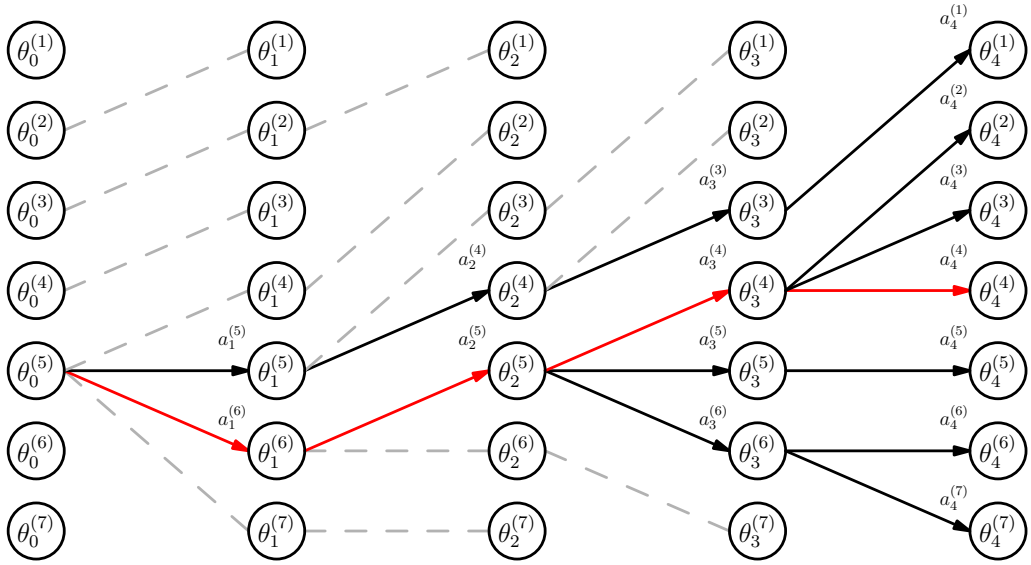


Figure 13.2: Trajectories in a general particle filter (solid lines) with a sampled trajectory (red line).

transitioned particle $\boldsymbol{\theta}_{t+1}^{(i)}$. If no resampling occurred, the mapping will be $a_{t+1}^{(i)} = i$, and if resampling occurred, some particles will probably share "parent" particles. A direct lineage from the last set of particles at $t = T$ to the original set of particles at $t = 0$ is usually called a *trajectory*. We will define the complete set of variables produced by the particle filter, that is, the set of particle trajectories, $\left\{\boldsymbol{\theta}_{0:T}^{(i)}\right\}_{i=1}^{N_p}$ along with its respective weights $\left\{w_{0:T}^{(i)}\right\}_{i=1}^{N_p}$ and particle ancestors, $\left\{a_{1:T}^{(i)}\right\}_{i=1}^{N_p}$ as the output of the particle filter. An illustration of the "surviving" trajectories is presented in Figure 13.2. It is important to note that within the context of PMCMC methods, the resampling strategy applied must satisfy the unbiasedness criterion as specified in Section 7.3 on page 87. As detailed in Andrieu *et al.* (2009), we can consider the resampling operation as the method to which particles at time t "choose" their ancestors at time $t-1$ as function of the parents' weights w_{t-1} as defined in (7.4), that is with probability

$$\mathcal{R}\left(\cdot|\left\{w_{t-1}^{(i)}\right\}_{i=1}^{N_p}\right).$$

The joint distribution of these variables, conditioned on a set of parameters $\boldsymbol{\Phi}$, can be denoted as

$$\begin{aligned}\psi_{\boldsymbol{\Phi}} &= \mathcal{PF}\left(\mathcal{D}_T, \boldsymbol{\Phi}\right) &(13.5)\\ &= \left\{\boldsymbol{\theta}_{0:T}^{(i)}, a_{1:T}^{(i)}, w_{0:T}^{(i)}\right\}_{i=1}^{N_p}, &(13.6)\end{aligned}$$

and our target distribution, in this case will be

$$q\left(\boldsymbol{\theta}_{0:T}\right) = p\left(\boldsymbol{\theta}_{0:T}|\mathcal{D}_T, \boldsymbol{\Phi}\right). \tag{13.7}$$

Algorithm 13.2 Particle Independent Metropolis-Hastings

initialise $n = 1$

run a particle filter, sampling $\boldsymbol{\theta}_{0:T}^{(1)} \sim p\left(\boldsymbol{\theta}_{0:T}|\mathcal{D}_T, \boldsymbol{\Phi}\right)$

calculate the likelihood $\widehat{\ell}^{(1)} = \prod_{t=1}^{T} \frac{1}{N_p} \sum_{i=1}^{N_p} w_t^{(i)}$

for $n > 1$

run a particle filter, sampling $\boldsymbol{\theta}_{0:T}^{\star} \sim p\left(\boldsymbol{\theta}_{0:T}|\mathcal{D}_T, \boldsymbol{\Phi}\right)$

calculate the likelihood $\widehat{\ell}^{\star} = \prod_{t=1}^{T} \frac{1}{N_p} \sum_{i=1}^{N_p} w_t^{(i)}$

calculate the probability

$$A = 1 \wedge \frac{\widehat{\ell}^{\star}}{\widehat{\ell}^{(n)}}$$

with probability A

accept $\boldsymbol{\theta}_{0:T}^{(n)} = \boldsymbol{\theta}_{0:T}^{\star}$ and $\widehat{\ell}^{(n)} = \widehat{\ell}^{\star}$

or reject $\boldsymbol{\theta}_{0:T}^{(n)} = \boldsymbol{\theta}_{0:T}^{(n-1)}$ and $\widehat{\ell}^{(n)} = \widehat{\ell}^{(n-1)}$

In this case, since it is not possible to sample directly from (13.7), the sample from an *extended proposal* distribution which will include all the variables produced by the SMC approximation, is denoted as $\psi_{\boldsymbol{\Phi}}$ in (13.5). Still according to Andrieu *et al.* (2009, 2010), by sampling an index k from the particle filter's last stage weights with probability $\{w_T\}_{i=1}^{N_p}$ (as illustrated in Figure 13.2) and the corresponding particle $\boldsymbol{\theta}_T^{(k)}$ and denoting the set of $\boldsymbol{\Theta}^N$ simulated values as

$$\overline{\boldsymbol{\theta}}_t = \left(\boldsymbol{\theta}_t^1, \boldsymbol{\theta}_t^2, \ldots, \boldsymbol{\theta}_t^N\right) \in \boldsymbol{\Theta}^N,$$

then the joint distribution of all particle filter generated values in (13.5) will be

$$q\left(k, \overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2, \ldots, \overline{\boldsymbol{\theta}}_T, a_1, a_2, \ldots, , a_T\right) = w_T^{(k)} \psi\left(\overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2, \ldots, \overline{\boldsymbol{\theta}}_T, a_1, a_2, \ldots, , a_T\right).$$

Andrieu *et al.* (2009, 2010) proves that by following this procedure we are producing a sample from the proposal (13.7), such that the PIMH target is $\pi\left(k, \overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2, \ldots, \overline{\boldsymbol{\theta}}_T, a_1, a_2, \ldots, , a_T\right)$ with a proposal $q\left(k, \overline{\boldsymbol{\theta}}_1, \overline{\boldsymbol{\theta}}_2, \ldots, \overline{\boldsymbol{\theta}}_T, a_1, a_2, \ldots, , a_T\right)$, therefore allowing us to calculate the acceptance ratio in Algorithm 13.2. If we consider that as the number of particles in the SMC estimator $N_p \to \infty$, so will $\widehat{p}(\boldsymbol{\theta}_{0:T}|\mathcal{D}_T, \boldsymbol{\Phi}) \to p\left(\boldsymbol{\theta}_{0:T}|\mathcal{D}_T, \boldsymbol{\Phi}\right)$ as consequently the acceptance ratio $A \to 1$.

We are, however, interested in the estimation of the joint posterior of states and parameters, $p(\boldsymbol{\Phi}, \boldsymbol{\theta}_{0:T}|\mathcal{D}_T)$ for which a particular method called Particle Marginal Metropolis-Hastings (PMMH) can be used. Since we know that the target distribution can be factored as

$$p(\boldsymbol{\Phi}, \boldsymbol{\theta}_{0:T}|\mathcal{D}_T) = p(\boldsymbol{\Phi}|\mathcal{D}_T)\, p(\boldsymbol{\theta}_{0:T}|\mathcal{D}_T, \boldsymbol{\Phi}),$$

our MH proposal distribution could take the form

$$q(\boldsymbol{\Phi}^\star, \boldsymbol{\theta}^\star_{0:T}) = q(\boldsymbol{\Phi}^\star|\boldsymbol{\Phi})\, p(\boldsymbol{\theta}^\star_{0:T}|\mathcal{D}_T, \boldsymbol{\Phi}^\star).$$

In this case, we would have a similar procedure to the PIMH, but now with an acceptance ratio (for the idealised scheme) of

$$\begin{aligned} A &= 1 \wedge \frac{p(\boldsymbol{\Phi}^\star, \boldsymbol{\theta}^\star_{0:T}|\mathcal{D}_T)\, q(\boldsymbol{\Phi}, \boldsymbol{\theta}^\star_{0:T}|\boldsymbol{\Phi}^\star, \mathcal{D}_T)}{p(\boldsymbol{\Phi}, \boldsymbol{\theta}^\star_{0:T}|\mathcal{D}_T)\, q(\boldsymbol{\Phi}^\star, \boldsymbol{\theta}^\star_{0:T}|\boldsymbol{\Phi}, \mathcal{D}_T)} \\ &= 1 \wedge \frac{p(\mathcal{D}_T|\boldsymbol{\Phi}^\star)\, q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^\star)}{p(\mathcal{D}_T|\boldsymbol{\Phi})\, q(\boldsymbol{\Phi}^\star|\boldsymbol{\Phi})}. \end{aligned}$$

In Andrieu *et al.* (2010) it is proven that PMMH leads to convergent algorithms and a notable property is that it leaves the distribution $p(\boldsymbol{\Phi}, \boldsymbol{\theta}_{0:T}|\mathcal{D}_T)$ invariant. Although theoretically the PMCMC will provide an exact estimation of the target distribution for any unbiased likelihood estimator with a number of particles $N_p \geq 1$ (Andrieu *et al.* (2009)), its efficiency will depend on the variance of the estimator. A higher number of particles could result in a better chain mixing, since in theory the variance of the likelihood estimator will grow linearly[2] with T. The choice of an appropriate number of particles N_p for the particle filter and also the scaling of the innovation variance in the random walk proposal are then very important aspects with a direct impact in the efficiency of the PMMH method. In Sherlock *et al.* (2015) it is advised that N_p should be such that the variance of the log-posterior is approximately 3. This can be calculated by performing a PMMH test run in order to calculate the variance of the log-posterior and selecting the number of particles accordingly. Still in Sherlock *et al.* (2015), the scaling, γ, of the innovation variance should be $\gamma = (2.38^2/d)$, where d is the dimension of the target. A general algorithm for PMMH is presented in Algorithm 13.3 on the next page.

Another method worth mentioning is Monte Carlo within Metropolis (MCWM), an alternative (approximate) version of PMMH. MCWM follows in essence the same approach as PMMH with the major difference that at each iteration, instead of keeping the marginal likelihood estimation from iteration $n-1$, that is $\ell^{(n-1)} = p\left(\mathcal{D}_T|\boldsymbol{\Phi}^{(n-1)}\right)$, this likelihood

[2] suggesting the number of particles for the particle filter estimation, N_p, should also grow linearly with T.

is recalculated independently along with $\ell^\star$. That is, at each iteration n, we calculate both

$$\begin{aligned} \ell^{(n-1)} &= \hat{p}\left(\mathcal{D}_T|\boldsymbol{\Phi}^{(n-1)}\right) \\ \ell^\star &= \hat{p}\left(\mathcal{D}_T|\boldsymbol{\Phi}^\star\right). \end{aligned}$$

As mentioned in Andrieu & Roberts (2009), although the invariant distribution of MCWM is not $\pi(\Phi)$, the samples generated by this method will be asymptotically distributed according to an approximation of $\pi(\Phi)$ increasing in precision as the number of particles N_p increase. A general algorithm for MCWM is presented in Algorithm 13.4 on the following page.

Algorithm 13.3 Particle Marginal Metropolis-Hastings

Initialise $n = 0$

sample $\boldsymbol{\Phi}^{(0)} \sim p(\boldsymbol{\Phi})$

run a particle filter to estimate $p\left(\boldsymbol{\theta}_{0:T}|\mathcal{D}_T, \boldsymbol{\Phi}^{(0)}\right)$

sample $\boldsymbol{\theta}_{0:T}^{(0)} \sim \widehat{p}\left(\cdot|\mathcal{D}_T, \boldsymbol{\Phi}^{(0)}\right)$

calculate $\ell^{(0)} = \widehat{p}\left(\mathcal{D}_T|\boldsymbol{\Phi}^{(0)}\right)$

Iteration $n \geq 1$

sample $\boldsymbol{\Phi}' \sim q\left(\cdot|\boldsymbol{\Phi}^{(n-1)}\right)$

run a particle filter

sample $\boldsymbol{\theta}'_{0:T} \sim p\left(\cdot|\mathcal{D}_T, \boldsymbol{\Phi}'\right)$

set $\ell' = \widehat{p}\left(\mathcal{D}_T|\boldsymbol{\Phi}'\right)$

calculate Metropolis-Hastings ratio

$$1 \wedge \frac{\ell' p\left(\boldsymbol{\Phi}'\right) q\left(\boldsymbol{\Phi}^{(n-1)}|\boldsymbol{\Phi}'\right)}{\ell^{(n-1)} p\left(\boldsymbol{\Phi}^{(n-1)}\right) q\left(\boldsymbol{\Phi}'|\boldsymbol{\Phi}^{(n-1)}\right)}$$

accept $\left\{\boldsymbol{\theta}_{0:T}^{(n)}, \boldsymbol{\Phi}^{(n)}\right\} = \left\{\boldsymbol{\theta}'_{0:T}, \boldsymbol{\Phi}'\right\}$, $\ell^{(n)} = \ell'$.

or reject $\left\{\boldsymbol{\theta}_{0:T}^{(n)}, \boldsymbol{\Phi}^{(n)}\right\} = \left\{\boldsymbol{\theta}_{0:T}^{(n-1)}, \boldsymbol{\Phi}^{(n-1)}\right\}$, $\ell^{(n)} = \ell^{(n-1)}$.

Example. Poisson AR(1) DGLM

Algorithm 13.4 Monte Carlo within Metropolis

Initialise $n = 0$

sample $\boldsymbol{\Phi}^{(0)} \sim p(\boldsymbol{\Phi})$

Iteration $n \geq 1$

sample $\boldsymbol{\Phi}' \sim q\left(\cdot|\boldsymbol{\Phi}^{(n-1)}\right)$

run a particle filter

sample $\boldsymbol{\theta}'_{0:T} \sim p(\cdot|\mathcal{D}_T, \boldsymbol{\Phi}')$

set $\ell' = \widehat{p}(\mathcal{D}_T|\boldsymbol{\Phi}')$

sample $\boldsymbol{\theta}_{0:T} \sim p\left(\cdot|\mathcal{D}_T, \boldsymbol{\Phi}^{(n-1)}\right)$

set $\ell^{(n-1)} = \widehat{p}\left(\mathcal{D}_T|\boldsymbol{\Phi}^{(n-1)}\right)$

calculate Metropolis-Hastings ratio

$$1 \wedge \frac{\ell' p(\boldsymbol{\Phi}') \, q\left(\boldsymbol{\Phi}^{(n-1)}|\boldsymbol{\Phi}'\right)}{\ell^{(n-1)} p\left(\boldsymbol{\Phi}^{(n-1)}\right) q\left(\boldsymbol{\Phi}'|\boldsymbol{\Phi}^{(n-1)}\right)}$$

accept $\left\{\boldsymbol{\theta}^{(n)}_{0:T}, \boldsymbol{\Phi}^{(n)}\right\} = \left\{\boldsymbol{\theta}'_{0:T}, \boldsymbol{\Phi}'\right\}$, $\ell^{(n)} = \ell'$.

or reject $\left\{\boldsymbol{\theta}^{(n)}_{0:T}, \boldsymbol{\Phi}^{(n)}\right\} = \left\{\boldsymbol{\theta}^{(n-1)}_{0:T}, \boldsymbol{\Phi}^{(n-1)}\right\}$, $\ell^{(n)} = \ell^{(n-1)}$.

The Autoregressive of order 1 (AR(1)) Poisson DLM can be specified as

$$\begin{aligned} y_t|\theta_t &\sim \text{Po}(\lambda_t) \\ \lambda_t &= \exp\{\theta_t\} \\ \theta_t|\theta_{t-1}, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\alpha + \beta\theta_{t-1}, \tau^2\right) \end{aligned}$$

where α and β are regression coefficients, in this case considered static but unknown. The parameter set used for the realisation of the $N_{obs} = 2000$ observations seen in Figure 13.3 was $\boldsymbol{\Phi} = \left\{\tau^2, \alpha, \beta\right\} = \{1.5, 0.7, 0.3\}$. For this model we can implement PMMH using a standard bootstrap filter (*i.e.* an SIR filter using the model's transition $p(\theta|\theta_{t-1})$ as the importance density) with $N_p = 3000$ to approximate the likelihood. The proposal step used was

$$\text{diag}(\boldsymbol{\Sigma}) = (0.1, 0.1, 0.1).$$

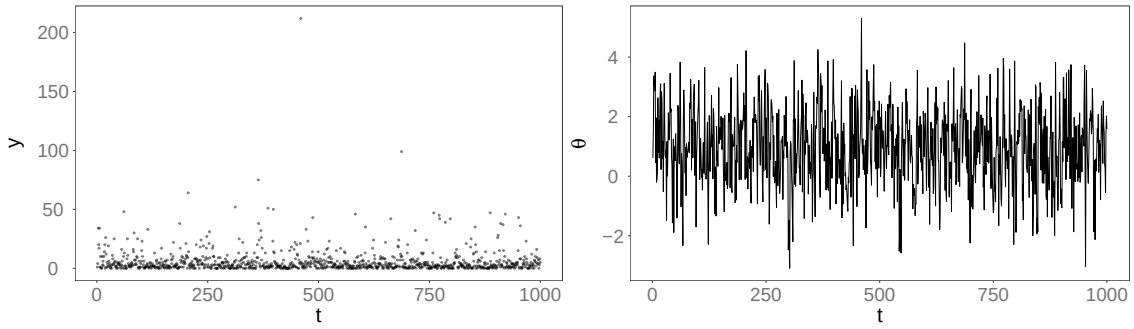


Figure 13.3: Observations $y_{1:N_{obs}}$ and states $\theta_{0:N_{obs}}$ for a realisation of a AR(1) PoDLM with $\boldsymbol{\Phi} = \{\tau^2, \alpha, \beta\} = \{1.5, 0.7, 0.3\}$.

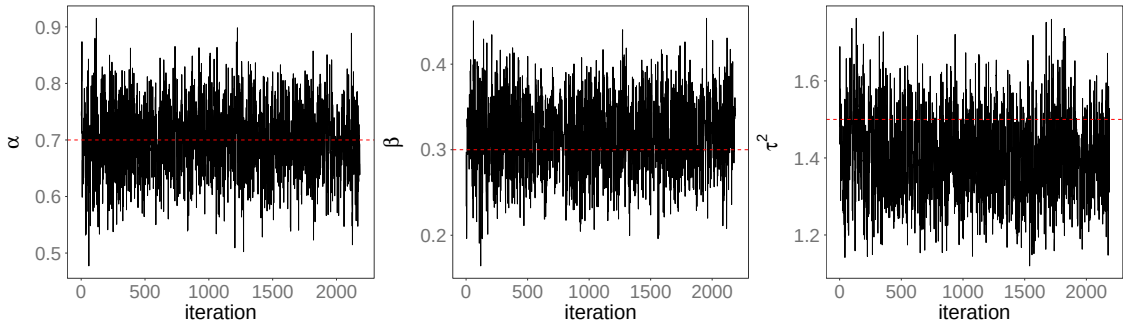
The priors used were

$$\begin{aligned} \tau_0^2 &\sim \mathcal{IG}(1,1) \\ \alpha_0 &\sim \text{Beta}(1,1) \\ \beta_0 &\sim \text{Beta}(1,1). \end{aligned}$$

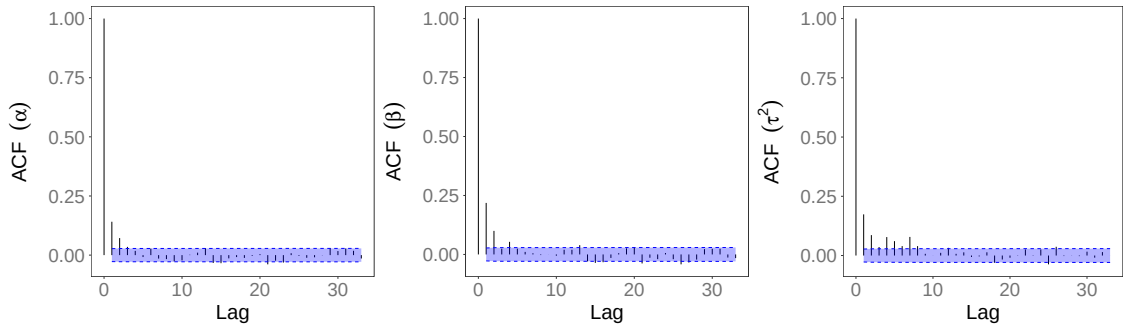
Represented in Figure 13.3 are observations, $y_{1:T}$, and states, $\theta_{0:T}$, for a realisation of the above model.

In Figure 13.4 we can see the traces for the parameter set $\boldsymbol{\Phi} = \left\{\alpha, \beta, \tau^2\right\}$ corresponding to $n = 219595$ iterations of the PMMH sampler. The burn-in period discarded was of 1000 samples after which a thinning factor of 1000 was applied, resulting in a final set of 2186 samples. The MCESS for the parameter set $\boldsymbol{\Phi} = \left\{\alpha, \beta, \tau^2\right\}$ was respectively $MCESS_{\boldsymbol{\Phi}} = \{1477.8, 1161.4, 925.9\}$.

For comparison, MCWM was applied to the same data, using the same parameter priors as for the PMMH. The same likelihood estimator was used (a SIR particle filter with $N_p = 3000$ and sampling from the prior as a proposal density) as well as the proposal step. The MCWM traces and k-lag ACF plots are presented in Figure 13.8 and a marginal density for $\boldsymbol{\Phi}$ at time $t = T$ is presented in Figure 13.9. A comparison of the latent state estimation for both PMMH and MCWM is presented in Figure 13.7. We can see from Figure 13.9 that the estimation of the parameters' posterior is consistent between both methods in this case and from Figure 13.7 we can see that the mean state's estimation is practically identical.



(a) Traces



(b) k-lag ACF for $\boldsymbol{\Phi} = \{\alpha, \beta, \tau^2\}$

Figure 13.4: AR(1) PoDLM PMMH traces (*top*) for $\boldsymbol{\Phi} = \{\alpha, \beta, \tau^2\}$. True parameters as the red horizontal line. k-lag ACF plots (*bottom*)

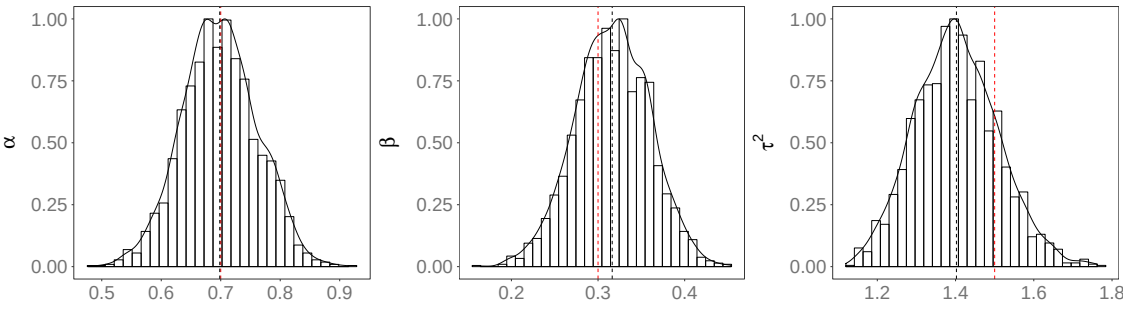


Figure 13.5: AR(1) PoDLM PMMH parameter posterior densities (normalised to 1) for $\boldsymbol{\Phi} = \{\alpha, \beta, \tau^2\}$. True parameters as the red vertical line.

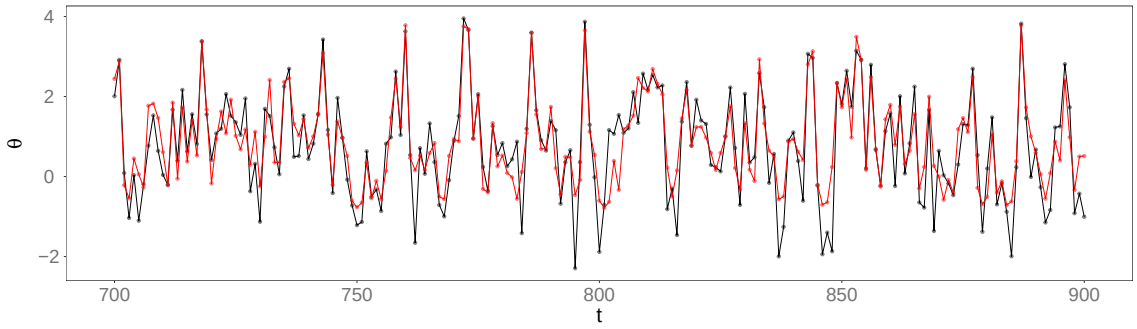


Figure 13.6: AR(1) PoDLM PMMH state posterior mean estimation (*red*) and true state from the realisation (*black*).

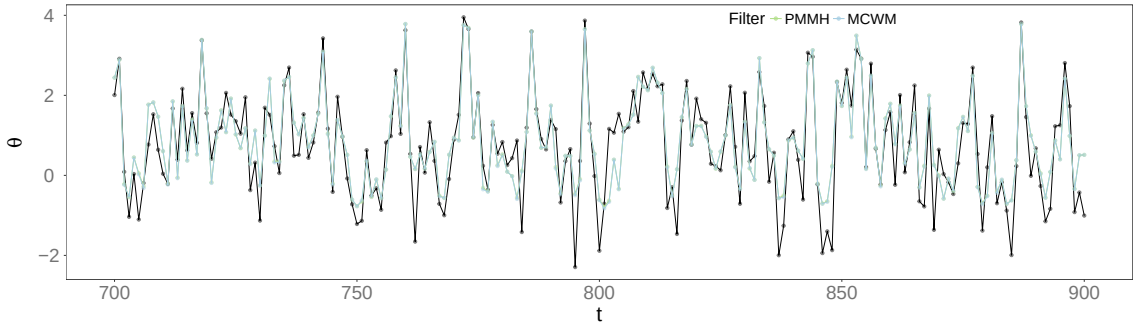
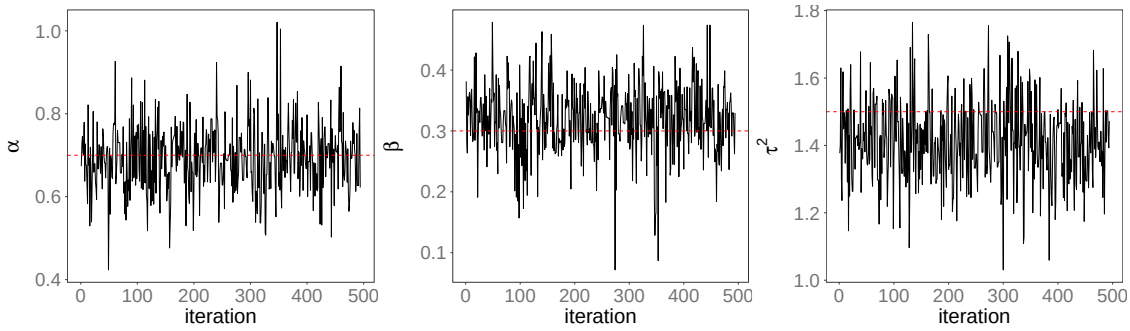
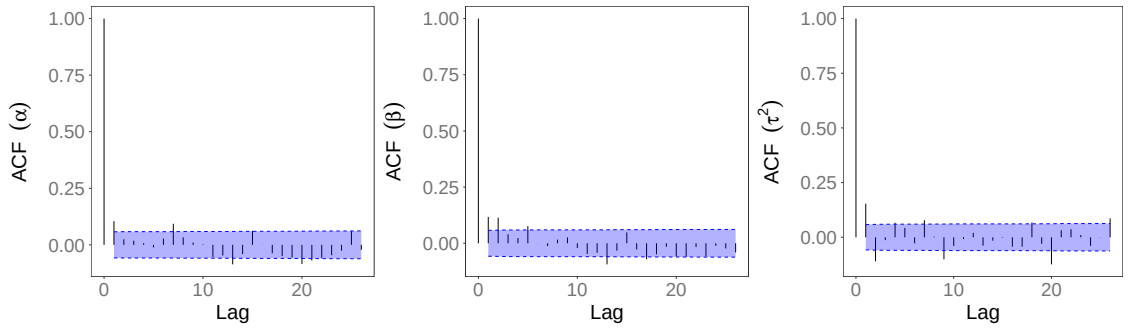


Figure 13.7: AR(1) PoDLM PMMH and MCWM state posterior mean estimation and true state from the realisation (*black*).



(a) Traces



(b) k-lag ACF for $\boldsymbol{\Phi} = \{\alpha, \beta, \tau^2\}$

Figure 13.8: AR(1) Poisson DGLM MCWM traces (*top*) for $\boldsymbol{\Phi} = \{\alpha, \beta, \tau^2\}$. True parameters as the red horizontal line. k-lag ACF plots (*bottom*)

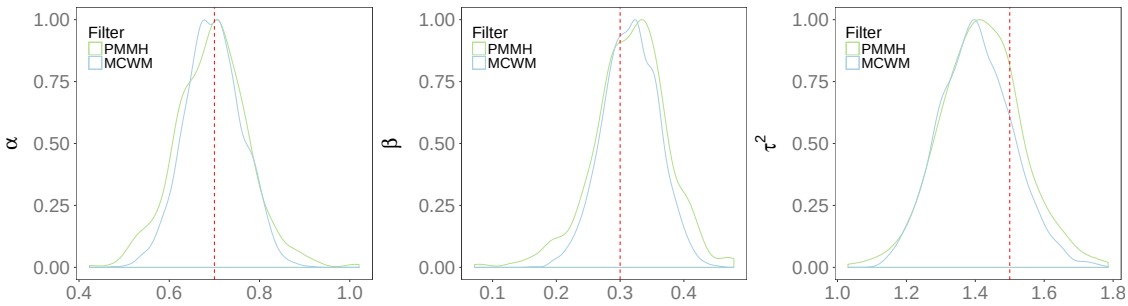


Figure 13.9: AR(1) Poisson DGLM PMMH and MCWM parameter posterior densities at $t = N_{obs}$ (normalised to 1) comparison for $\boldsymbol{\Phi} = \{\alpha, \beta, \tau^2\}$. True parameters as the red vertical line.

Chapter 14

Iterated Batch Importance Sampling

A method presented by Chopin (2002) to perform sequential (but not online) state and parameter estimation is the Iterated Batch Importance Sampling (IBIS) algorithm. For parameter estimation, IBIS targets the posterior $\pi\left(\boldsymbol{\Phi}|\mathcal{D}_t\right)$, provided some conditions are met regarding the actual model.

Chopin (2002) proposes that inference about the parameter posterior can be calculated in batches

$$p_k\left(\boldsymbol{\Phi}|\mathcal{D}_k\right), \qquad k = 1 < \cdots < t < \cdots < N_{obs},$$

where $p_k\left(\boldsymbol{\Phi}|\mathcal{D}_k\right)$ indicates the partial posterior up until $t = k$.

Chopin (2002) states that, considering the first batch of observations $\mathcal{D}_{k_1}$, using a particle system targeting to perform inference on $p_{k_1}\left(\boldsymbol{\Phi}|\mathcal{D}_{k_1}\right)$, we can assume that whenever a new batch of observations is available, $\mathcal{D}_{k_2}$ (with $k_2 > k_1$) then it is likely that

$$p_{k_1}\left(\boldsymbol{\Phi}|\mathcal{D}_{k_1}\right) \approx p_{k_2}\left(\boldsymbol{\Phi}|\mathcal{D}_{k_2}\right).$$

With this new batch of observations, importance sampling can then be applied to the particle system with importance weights defined by

$$w \propto \frac{p\left(\boldsymbol{\Phi}|\mathcal{D}_{k_2}\right)}{p\left(\boldsymbol{\Phi}|\mathcal{D}_{k_1}\right)} \propto \frac{p\left(\mathcal{D}_{k_2}|\boldsymbol{\Phi}\right)}{p\left(\mathcal{D}_{k_1}|\boldsymbol{\Phi}\right)} \propto p\left(y_{k_1+1:k_2}|y_{1:k_1}, \boldsymbol{\Phi}\right).$$

To minimise weight degeneracy, the current particle set is built on a resample-sample framework, where resample occurs whenever the ESS falls below a certain threshold. In

the case of IBIS the ESS is calculated as

$$\widehat{ESS} = \frac{\left(\sum_{i=1}^{N_{\Phi}} w_t^{(i)}\right)^2}{\sum_{i=1}^{N_{\Phi}} \left(w_t^{(i)}\right)^2}. \tag{14.1}$$

As seen in Section 7.1.1, resampling methods will help with degeneracy problems, but will introduce particle impoverishment. To address this problem, Chopin (2002); Gilks & Berzuini (2001) perform a particle rejuvenation step by use of a Markov kernel. If we consider a Markov kernel K_t with a stationary distribution, then at the sample stage (after resampling) the parameter particles will be simulated from

$$\boldsymbol{\Phi}_t^{(i)} \sim \frac{1}{\sum_{i=1}^{N_{\Phi}} w_t^{(i)}} \sum_{i=1}^{N_{\Phi}} w_t^{(i)} K_t\left(\boldsymbol{\Phi}^{(i)}, d\boldsymbol{\Phi}\right).$$

A common choice of K_t is a random-walk Metropolis-Hastings (MH) such as

$$\boldsymbol{\Phi}_t^{\star} \sim \mathcal{N}\left(\boldsymbol{\Phi}_t, \Sigma\right)$$

with Σ proportional to the $N_{\boldsymbol{\Phi}}$ parameter particles variance[1]

$$\Sigma = k\mathrm{Var}\left[\boldsymbol{\Phi}_t | \mathcal{D}_{t-1}\right].$$

One of the constraints is that given the parameter prior $p\left(\boldsymbol{\Phi}\right)$ and the observations $\mathcal{D}_t$ we are able to calculate the likelihood

$$p\left(y_t | \mathcal{D}_{t-1}, \boldsymbol{\Phi}\right). \tag{14.2}$$

If we are able to calculate (14.2), then a sequential importance sampling approach is taken where a discrete set of $N_{\boldsymbol{\Phi}}$ *parameter particles* will approximate the posterior $p\left(\boldsymbol{\Phi}|\mathcal{D}_t\right)$. To do so, we assume that at each time t, the approximation at time $t-1$, $\left\{\boldsymbol{\Phi}_{t-1}^{(i)}, w_{t-1}^{(i)}\right\}_{i=1}^{N_{\Phi}}$, is updated to $\left\{\boldsymbol{\Phi}_t^{(i)}, w_t^{(i)}\right\}_{i=1}^{N_{\Phi}}$ using importance weights proportional to the likelihood in (14.2):

$$\begin{aligned} w_t^{(i)} &\propto \frac{p\left(\boldsymbol{\Phi}|\mathcal{D}_t\right)}{p\left(\boldsymbol{\Phi}|\mathcal{D}_{t-1}\right)} \\ &\propto p\left(y_t|\mathcal{D}_{t-1}, \boldsymbol{\Phi}\right). \end{aligned}$$

As seen in previous sections, resampling introduces a particle impoverishment problem

[1] For a different approach, where the kernels can be chosen adaptively, *c.f.* Fearnhead & Taylor (2013).

and one of the main characteristics of IBIS, as noted in Chopin (2002), is the introduction of parameter particle rejuvenation. The new proposed parameters $\boldsymbol{\Phi}^{(i)\star}$ can then be accepted according to the acceptance ratio

$$A = \min \left\{ 1, \frac{p\left(\boldsymbol{\Phi}_t^{\star(i)}\right) p\left(\mathcal{D}_t|\boldsymbol{\Phi}_t^{\star(i)}\right) q\left(\boldsymbol{\Phi}_t^{(i)}|\boldsymbol{\Phi}_t^{\star(i)}\right)}{p\left(\boldsymbol{\Phi}_t^{(i)}\right) p\left(\mathcal{D}_t|\boldsymbol{\Phi}_t^{(i)}\right) q\left(\boldsymbol{\Phi}_t^{\star(i)}|\boldsymbol{\Phi}_t^{(i)}\right)} \right\}.$$

Given the $N_{\boldsymbol{\Phi}}$ particles, IBIS allows us (Chopin *et al.* (2013)) to know characteristics of the target distribution:

$$\begin{aligned} \widehat{\Sigma} &= \frac{1}{\sum_{i=1}^{N_\Phi} w^{(i)}} \sum_{i=1}^{N_\Phi} w^{(i)} \left(\boldsymbol{\Phi}^{(i)} - \widehat{\boldsymbol{\mu}}\right)\left(\boldsymbol{\Phi}^{(i)} - \widehat{\boldsymbol{\mu}}\right)^T \\ \widehat{\boldsymbol{\mu}} &= \frac{1}{\sum_{i=1}^{N_\Phi} w^{(i)}} \sum w^{(i)} \boldsymbol{\Phi}^{(i)}. \end{aligned}$$

Although the likelihood in (14.2) is typically intractable, it is clear that one model for which IBIS is directly applicable is the NDLM, since as we've seen from Section 4, the incremental likelihood (14.2) can be easily calculated.[2] As detailed in (4.9), this will be the KF's *predictive density*:

$$p\left(y_t|\mathcal{D}_{t-1}, \boldsymbol{\Phi}\right) = \mathcal{N}\left(y_t; f_t, Q_t\right).$$

In the NDLM case, the total data likelihood $p\left(\mathcal{D}_t|\boldsymbol{\Phi}\right)$, used in the MH acceptance ratio can also be easily calculated, since

$$\begin{aligned} p\left(\mathcal{D}_t|\boldsymbol{\Phi}\right) &= p\left(y_1|\boldsymbol{\Phi}\right) \prod_{k=2}^{t} p\left(y_k|\mathcal{D}_{k-1}, \boldsymbol{\Phi}\right) \\ &= \prod_{k=1}^{t} \mathcal{N}\left(f_k, Q_k\right). \end{aligned}$$

For this class of models, IBIS can then be implemented as a concurrent system of $N_{\boldsymbol{\Phi}}$ Kalman filters, each conditioned on a parameter $\left\{\boldsymbol{\Phi}^{(i)}\right\}_{i=1}^{N_{\boldsymbol{\Phi}}}$ applying the Kalman recursions of Section 4 to calculate $p\left(y_t|\mathcal{D}_{t-1}, \boldsymbol{\Phi}^{(i)}\right)$. At each time t, the set comprising the KFs first and second moments, along with the importance weights and the parameters:

$$\left\{\mathbf{m}_t^{(i)}, \mathsf{C}_t^{(i)}, w_t^{(i)}, \boldsymbol{\Phi}_t^{(i)}\right\}_{i=1}^{N_\Phi},$$

[2] For the moment, for simplicity we will drop the particle notation (i), later generalising the results to a system of particles.

are then sufficient to implement the IBIS algorithm allowing for the MH ratio to be calculated and new parameter proposed.

As an example we will look at a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM model, that is taking the form

$$\begin{aligned} y_t|\theta_t, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\theta_t, \nu^2\right) \\ \theta_t|\theta_{t-1}, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\theta_{t-1}, \tau^2\right). \end{aligned}$$

The state priors used were $m_0 = 0$ and $C_0 = 100$. The parameters for the realisation (Figure 14.1) are $\nu^2 = 4.2$ and $\tau^2 = 2.3$ and the estimation was performed using $N_{\boldsymbol{\Phi}} = 10000$. Assuming we have parameter priors of $p\left(\tau_0^2\right) = \mathcal{IG}(1,1)$ and $p\left(\nu_0^2\right) = \mathcal{IG}(1,1)$, at each time point t we calculate the new update KF moments according to (4.11) and (4.12) for each particle

$$\begin{aligned} m_t^{(i)} &= m_{t-1}^{(i)} + \frac{R_t^{(i)}}{R_t^{(i)} + \nu_t^{2(i)}}\left(y_t - m_{t-1}^{(i)}\right) \\ C_t^{(i)} &= R_t^{(i)} - \frac{R_t^{2(i)}}{R_t^{(i)} + \tau_t^{2(i)}} \\ R_t^{(i)} &= C_{t-1}^{(i)} + \tau_t^{2(i)}. \end{aligned}$$

At time point t the importance weights are calculated using the intermediate quantities f_t and Q_t such that

$$w_t^{(i)} \propto \mathcal{N}\left(\underbrace{m_{t-1}^{(i)}}_{f_t}, \underbrace{C_{t-1}^{(i)} + \tau_t^{2(i)} + \nu_t^{2(i)}}_{Q_t}\right),$$

and the $\widehat{ESS}$ calculated according to (14.1). If at any point the $\widehat{ESS}$ falls below the predefined threshold (in this case $\widehat{ESS}_{eff} = \frac{N_p}{2}$), the particles are resampled (using multinomial resampling) according to $w_t^{(i)}$ and new parameters are proposed from a rejuvenation kernel K_t such that

$$\begin{aligned} \boldsymbol{\Phi}_t^{\star(i)} &\sim \mathcal{N}\left(\boldsymbol{\Phi}_t^{(i)}, \Sigma\right) \\ \Sigma &= \mathrm{Var}\left[\boldsymbol{\Phi}_t^{(i)}|\mathcal{D}_t\right]. \end{aligned}$$

The MH ratio is given by

$$A_t = \min\left\{1, \frac{p\left(\boldsymbol{\Phi}_t^{\star(i)}\right) q\left(\boldsymbol{\Phi}_t^{(i)}|\boldsymbol{\Phi}_t^{\star(i)}\right) L_t^{\star}}{p\left(\boldsymbol{\Phi}_t^{(i)}\right) q\left(\boldsymbol{\Phi}_t^{\star(i)}|\boldsymbol{\Phi}_t^{(i)}\right) L_t}\right\} \tag{14.3}$$

where

$$L_t^\star = \prod_{k=1}^{t} \mathcal{N}\left(y_k; f_k^\star, Q_k^\star\right)$$
$$L_t = \prod_{k=1}^{t} \mathcal{N}\left(y_k; f_k, Q_k\right),$$

and where $\{f_k^\star, Q_k^\star\}_{k=1}^t$ and $\{f_k, Q_k\}_{k=1}^t$ refer to the predictive density moments calculated using respectively the parameter set $\boldsymbol{\Phi}_t^{\star(i)}$ and $\boldsymbol{\Phi}_t^{(i)}$, using a batch size of $p = 1$. If we were interested simply in parameter inference, it would suffice to keep a running total of the marginal likelihoods. However, in the context of this thesis, since we are interested in both state and parameter inference, it is necessary to keep the entire history of particles

$$\left\{ \left\{ \mathbf{m}_0^{(i)}, \mathsf{C}_0^{(i)}, \boldsymbol{\Phi}_0^{(i)} \right\}_{i=1}^{N_\Phi}, \left\{ \mathbf{m}_1^{(i)}, \mathsf{C}_1^{(i)}, \boldsymbol{\Phi}_1^{(i)} \right\}_{i=1}^{N_\Phi}, \cdots, \left\{ \mathbf{m}_t^{(i)}, \mathsf{C}_t^{(i)}, \boldsymbol{\Phi}_t^{(i)} \right\}_{i=1}^{N_\Phi} \right\}.$$

An illustration of the estimation history for τ^2 and ν^2 is given in Figures 14.2a and 14.2b with vertical dotted lines indicating the time points where resampling occurred.

The fact that the entire history quantities must be stored for the rejuvenation step is the reason why IBIS, although a sequential method, cannot be considered an online one. The computation times will grow with t making it infeasible to use it for long running estimations. However, as stated in Chopin (2002), assuming a fixed degeneracy criterion and that the efficiency of the move step is constant, the number of new points needed for each new batch of observations will increase geometrically. This simulation performed is consistent in this result as we can notice a decrease in frequency for the resampling stages in Figure 14.3 (dotted lines). Regarding the acceptances of MH, we can also see from Figure 14.4, that as expected, since we started from an uninformative prior, they fluctuate considerably for the initial moves, but as the posterior $p_k\left(\boldsymbol{\Phi}|\mathcal{D}_k\right)$ tends to $p\left(\boldsymbol{\Phi}|\mathcal{D}_T\right)$ they do stabilise in a consistent manner with the decrease of frequency of resampling stages.

This is clear from Figure 14.3 where we can see the cumulative running time of the algorithm displaying a clearly non-linear increase, with considerable delays at the total resampling points, where the total likelihood calculation occurs. In an online method, we would expect to see a definite linear time increase regardless of the data size.

As mentioned previously, it is clear that one of the key requirements of the IBIS algorithm is that we are able to calculate the *incremental likelihood* $p\left(y_{(t+1):(t+p)}|y_{1:t}, \boldsymbol{\Phi}\right)$. Although this is possible for certain models (namely the Gaussian DLM) it may not be possible for other types of DGLM. For those models, an alternate method, SMC^2, built on the concepts of IBIS is available and detailed in Chapter 15. For the remainder of this thesis, especially in Part V we will use the special case where the batch size is set to $p = 1$.

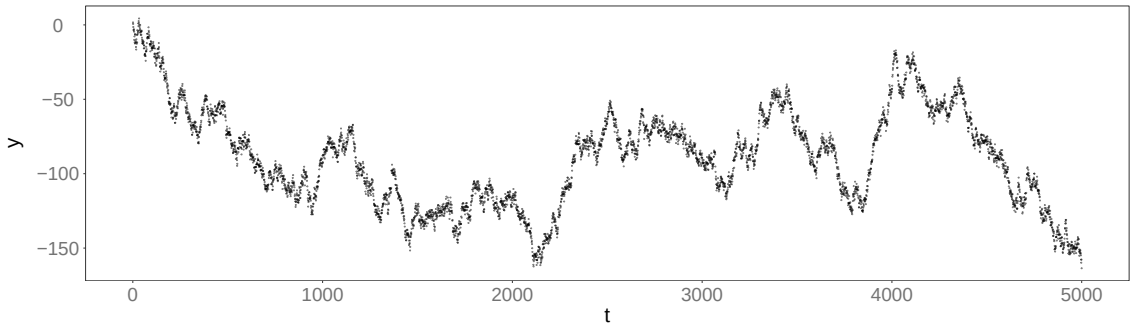
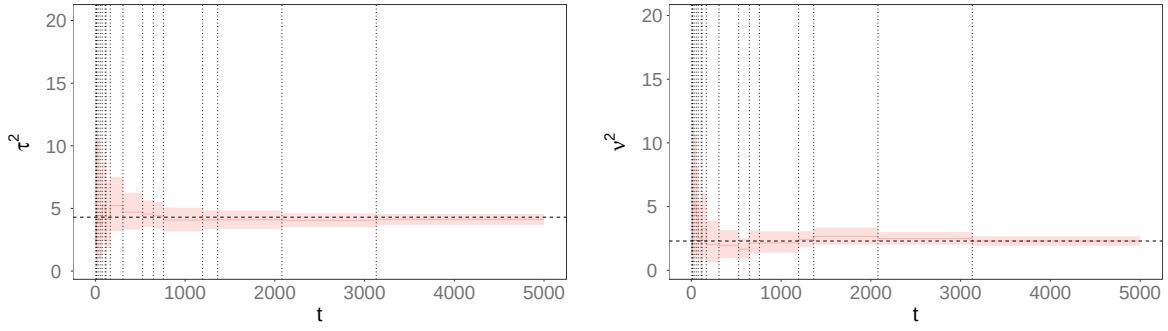
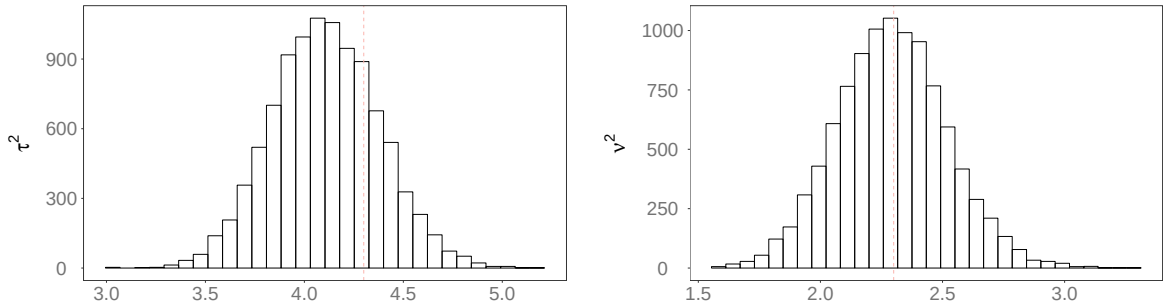


Figure 14.1: Observations from a realisation of a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM with $\boldsymbol{\Phi} = \left\{\nu^2, \tau^2\right\} = \{4.2, 2.3\}$.



(a) $\tau^2_{0:T}$ (b) $\nu^2_{0:T}$



(c) $p\left(\tau^2|\mathcal{D}_T\right)$ (d) $p\left(\nu^2|\mathcal{D}_T\right)$

Figure 14.2: Parameter posterior estimation history for τ^2 and ν^2 (colour line represents posterior mean, shaded area 90% equitailed credibility intervaland vertical line the rejuvenation stages) and (normalised) parameter posteriors at $t = T$ for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM using IBIS with $N_{\boldsymbol{\Phi}} = 10000$. Dashed vertical red lines represent the true values.

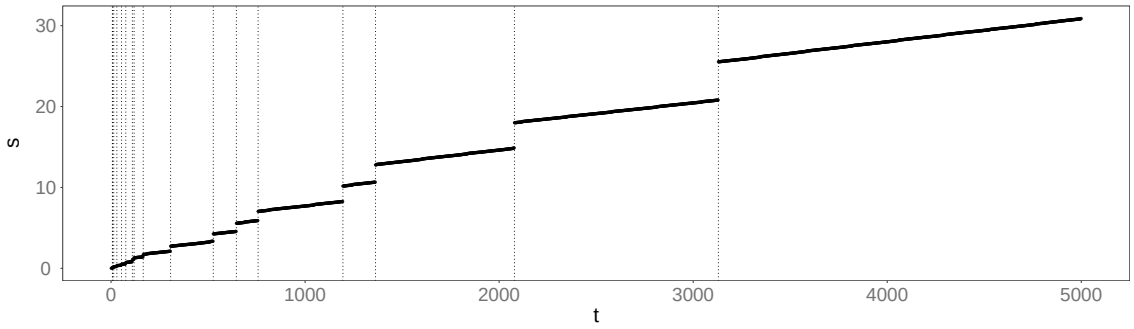


Figure 14.3: Cumulative computational time for IBIS. Vertical lines represent the resampling stages.

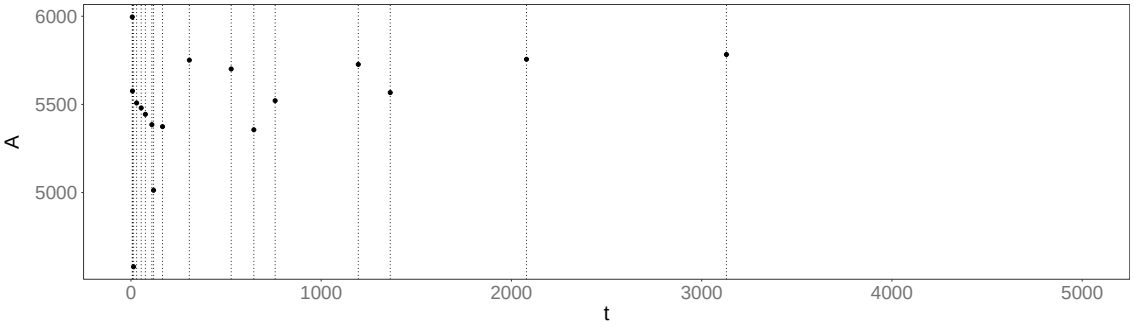


Figure 14.4: Number of Metropolis-Hastings acceptances for rejuvenation step of IBIS with $N_{\boldsymbol{\Phi}} = 10000$. Vertical lines represent the resampling steps.

Since the KF recursions play such a crucial role in the IBIS method, we will also implement a variant of IBIS using the SVD implementation of the KF (as described in Section 4.1).

14.1 Online IBIS

A possible solution to use the rejuvenation method of IBIS but in an online fashion is to use a sliding window of observations, thus approximating the total likelihood by a partial likelihood of the data. If assume we that

$$p\left(\boldsymbol{\Phi}|\mathcal{D}_t\right) \approx p\left(\boldsymbol{\Phi}|y_{(t-k):t}\right)$$

for a sufficiently large window $k > 0$, we will not be targeting the correct posterior $p\left(\boldsymbol{\Phi}|\mathcal{D}_t\right)$ but a good approximation $p\left(\tilde{\boldsymbol{\Phi}}|y_{t-k:t}\right)$. By discarding observations in the far horizon, this will in essence amount to the estimation of an artificially dynamic parameter $\tilde{\boldsymbol{\Phi}}_t$ that would be close to $\boldsymbol{\Phi}$. Not having to store the entirety of the information, but only a subset, allows us to perform online estimation. Although, depending on the window size

Algorithm 14.1 IBIS

initialisation

draw $\left\{\boldsymbol{\Phi}_0^{(i)}, w_0^{(i)}\right\}_{i=1}^{N_{\Phi}} \sim \pi\left(\boldsymbol{\Phi}_0\right)$

set N_{eff}

for $t \leftarrow 1$ to N_{obs}

for $i \leftarrow 1$ to $N_{\boldsymbol{\Phi}}$

calculate the weights according to

$$w_t^{(i)} \propto w_{t-1}^{(i)} p\left(y_{(t+1):(t+p)} | \mathcal{D}_t, \boldsymbol{\Phi}_t^{(i)}\right)$$

if $\widehat{ESS} < N_{eff}$

resample $\left\{\boldsymbol{\Phi}_t^{(i)}, w_t^{(i)}\right\}_{i=1}^{N_{\Phi}} \rightarrow \left\{\boldsymbol{\Phi}^{(k)}, 1\right\}_{i=1}^{N_{\Phi}}$

draw

$$\boldsymbol{\Phi}_{t+p}^{(i)} \sim K_{t+p}\left(\boldsymbol{\Phi}^{(i)}, \cdot\right), \qquad i = 1, \ldots, N_{\Phi},$$

where K_{n+p} is a transition kernel with stationary distribution π_{t+p}.

set $t = t + p$ and $\{\boldsymbol{\Phi}^{(i)}, w^{(i)}\}_{i=1}^{N_{\Phi}} = \{\boldsymbol{\Phi}^{(k)}, 1\}_{i=1}^{N_{\Phi}}$.

k there will be a computational impact at each resampling step, it will now at least be bounded and not growing along with the size of the data, effectively trading a computational cost at the resampling stage (for a constant number of particles $N_{\boldsymbol{\Phi}}$) of $\mathcal{O}(t)$ in full IBIS for $\mathcal{O}(1)$ in the online IBIS. The parameters will still be proposed from a random-walk MH kernel such that

$$\begin{aligned} \tilde{\boldsymbol{\Phi}}_t^{\star(i)} &\sim \mathcal{N}\left(\tilde{\boldsymbol{\Phi}}_t^{(i)}, \Sigma_t\right), \\ \Sigma_t &= \operatorname{Var}\left[\tilde{\boldsymbol{\Phi}}_t^{(i)} | y_{(t-k):t}\right]. \end{aligned}$$

and the acceptance ratio will be calculated as in (14.3), but now according the partial data likelihoods

$$\begin{aligned} \tilde{L}_t^{\star} &= \prod_{n=t-k}^{t} \mathcal{N}\left(y_n; f_n^{\star}, Q_n^{\star}\right) \\ \tilde{L}_t &= \prod_{n=t-k}^{t} \mathcal{N}\left(y_n; f_n, Q_n\right), \end{aligned}$$

where $\{f_n, Q_n\}_{n=t-k}^{t}$ and $\{f_n^\star, Q_n^\star\}_{n=t-k}^{t}$ refer to the predictive density moments calculated using respectively the parameter set $\tilde{\boldsymbol{\Phi}}_t^{(i)}$ and $\tilde{\boldsymbol{\Phi}}_t^{\star(i)}$.

Using the same model and data as in Section 14 on page 165, that is a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM taking the form

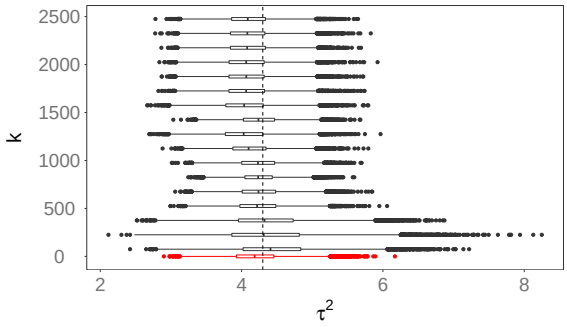
$$\begin{aligned} y_t|\theta_t, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\theta_t, \nu^2\right) \\ \theta_t|\theta_{t-1}, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\theta_{t-1}.\tau^2\right). \end{aligned}$$

The state priors used were $\theta_0 \sim \mathcal{N}(0, 100)$ and with parameters $\nu^2 = 4.2$ and $\tau^2 = 2.3$. The number of particles used where $N_p = 10^4$, the resampling algorithm was multinomial resampling and the resampling criteria was $\widehat{ESS}_{eff} = \frac{1}{2}N_p$. The priors used for the parameters were also the same as the example 14 on page 165, that is $p\left(\tau_0^2\right) = \mathcal{IG}(1,1)$ and $p\left(\nu_0^2\right) = \mathcal{IG}(1,1)$.

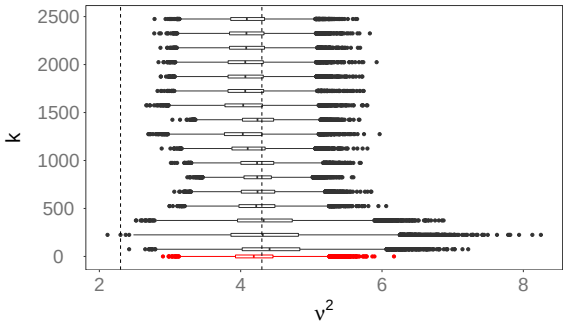
Using the nomenclature $\widetilde{\mathcal{D}}_T^k = \left\{y_{(T-k),} y_{(T-k+1),} \ldots, y_{T,}\right\}$, we can see in Figure 14.5 the $p\left(\tau^2|\widetilde{\mathcal{D}}_T^k\right)$ and $p\left(\nu^2|\widetilde{\mathcal{D}}_T^k\right)$ marginals using a range of window sizes k varying from 50 to $N_{obs}/2 = 2500$. Although, as expected, there is some variability we can see that the even for very small window sizes ($k \approx 50$ observations) the mean value of $\tilde{\boldsymbol{\Phi}}_t$ is consistent with the IBIS estimate using the entirety of the data (in red) although with some over estimation of the variance for small window sizes. It is worth noting that in the limit case where the window coincides with the totality of the observations, $k = N_{obs}$, O-IBIS is equivalent to IBIS. Since in this case O-IBIS will (as IBIS) target $p\left(\boldsymbol{\Phi}|\mathcal{D}_t^{N_{obs}}\right) = p(\boldsymbol{\Phi}|\mathcal{D}_t)$, we should expect that the posterior estimation should match. We can see in Figure 14.5 and Table 14.1, that with the value of $k > 1500$ the posterior estimation using O-IBIS stabilises as k increases, providing an approximation consistent with $p(\boldsymbol{\Phi}|\mathcal{D}_t)$ (using IBIS, in red).

One of the characteristics of IBIS is that the resampling steps will occur with less frequency the more data there is available to estimate the parameters. If we look at Figure 14.6a, showing the total computational time in seconds for the online IBIS with different k window sizes, we can see that the cumulative time slope remains constant approximately around $k = 1000$. Although it is clear that with higher window sizes we would expect (given the same number of resample-move steps) the computational cost to grow, we can see also from Figure 14.6b that for very small windows ($k \approx 50$) the resampling steps are much more frequent. After values of $k \approx 1000$, the number resampling steps diminishes and remains almost constant.

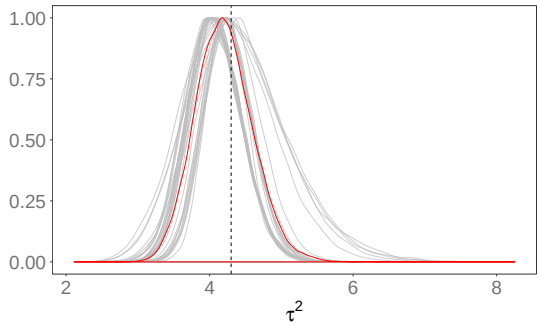
It is important to note that there is no practical advantage in using very small window sizes as it negatively impacts the quality of the approximation, but that by using a fixed number of past observations has as its main advantage to avoid the computational cost



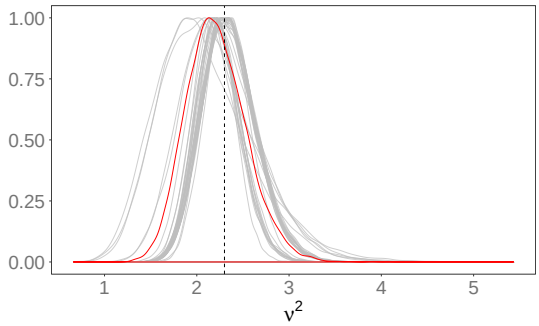
(a) $p\left(\tau^2|\widetilde{\mathcal{D}}_T^k\right)$ for different values of k using O-IBIS. Red line represents the parameter posterior at $t = T$ using IBIS.



(b) $p\left(\nu^2|\widetilde{\mathcal{D}}_T^k\right)$ for different values of k using O-IBIS. Red line represents the parameter posterior at $t = T$ using IBIS.



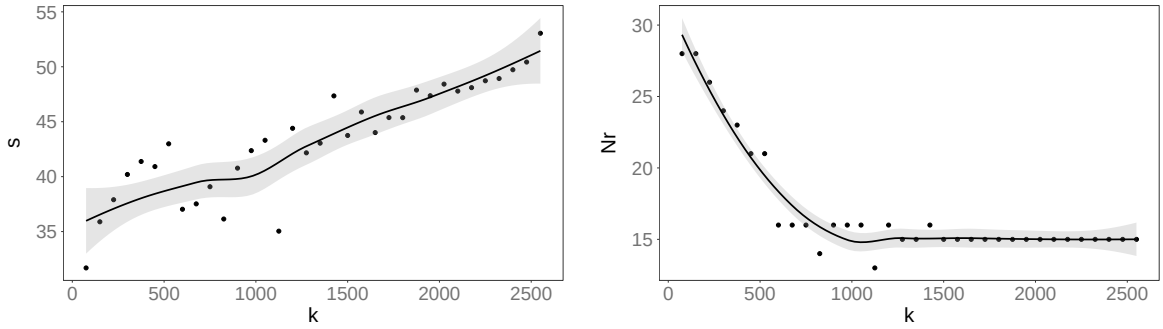
(c) $p\left(\tau^2|\widetilde{\mathcal{D}}_T^k\right)$ (normalised). Red line represents the parameter posterior at $t = T$ using IBIS.



(d) $p\left(\nu^2|\widetilde{\mathcal{D}}_T^k\right)$ (normalised). Red line represents the parameter posterior at $t = T$ using IBIS.

Figure 14.5: $p\left(\tau^2|\widetilde{\mathcal{D}}_T^k\right)$ and $p\left(\nu^2|\widetilde{\mathcal{D}}_T^k\right)$ marginals for a sliding window IBIS for different values of k. Estimation using IBIS with the entirety of the data ($k = T$) in red for comparison.

to grow alongside with the number of observations and allow an online inference on the model's parameters.



(a) Total computation time in seconds

(b) Number of resampling steps

Figure 14.6: Total computation time and number of resampling steps for a $\mathcal{M} = \{\mathcal{P}(1)\}$ NDLM for IBIS with a sliding window for the observations using different window sizes k.

k	$\overline{\tau}^2$	$\overline{\nu}^2$	k	$\overline{\tau}^2$	$\overline{\nu}^2$
75	4.4515 (0.6143)	2.2579 (0.4193)	1425	4.2512 (0.3262)	2.2323 (0.2325)
150	4.3874 (0.6872)	2.1121 (0.495)	1500	4.0436 (0.3919)	2.3379 (0.2947)
225	4.374 (0.7223)	2.1325 (0.5231)	1575	4.0476 (0.3893)	2.3368 (0.2938)
300	4.38 (0.68)	2.0477 (0.4961)	1650	4.0782 (0.3789)	2.3516 (0.2919)
375	4.3593 (0.5799)	2.2704 (0.4159)	1725	4.0814 (0.3743)	2.351 (0.2919)
450	4.33 (0.3956)	2.1793 (0.2705)	1800	4.0723 (0.3765)	2.3471 (0.2954)
525	4.2389 (0.3788)	2.2323 (0.2631)	1875	4.0786 (0.3711)	2.3485 (0.2967)
600	4.2632 (0.3505)	2.25 (0.241)	1950	4.0755 (0.3709)	2.35 (0.2866)
675	4.2526 (0.3492)	2.2504 (0.2426)	2025	4.0791 (0.3693)	2.3538 (0.2885)
750	4.2639 (0.3415)	2.2449 (0.2447)	2100	4.0912 (0.363)	2.3686 (0.2924)
825	4.2464 (0.2894)	2.2278 (0.1947)	2175	4.0952 (0.3627)	2.3652 (0.2959)
900	4.2711 (0.3453)	2.2314 (0.2437)	2250	4.0947 (0.3623)	2.3745 (0.2985)
975	4.2429 (0.3455)	2.2102 (0.2493)	2325	4.0919 (0.3611)	2.3663 (0.2922)
1050	4.2736 (0.3307)	2.2347 (0.2397)	2400	4.1003 (0.3556)	2.3728 (0.2946)
1125	4.1177 (0.3491)	2.3372 (0.2451)	2475	4.1007 (0.3557)	2.3771 (0.3013)
1200	4.27 (0.3314)	2.228 (0.2384)	2550	4.0991 (0.3536)	2.3672 (0.296)
1350	4.0491 (0.3872)	2.3429 (0.2955)			
IBIS	4.2018 (0.3972)	2.2134 (0.3254)	IBIS	4.2018 (0.3972)	2.2134 (0.3254)

Table 14.1: Summary of parameter posterior mean and standard deviation (in brackets) for τ^2 and ν^2 using O-IBIS with different window sizes k. IBIS posterior mean and standard deviation included for comparison.

Chapter 15

SMC^2

SMC^2, like IBIS, is a sequential (but not online) method introduced in Chopin *et al.* (2013), which targets the sequence of distributions

$$\left\{p\left(\boldsymbol{\Phi}\right), p\left(\boldsymbol{\Phi}|y_1\right), p\left(\boldsymbol{\Phi}|y_{1:2}\right), \ldots, p\left(\boldsymbol{\Phi}|\mathcal{D}_T\right)\right\}. \tag{15.1}$$

As seen in Section 14 on page 165, one of requirements of IBIS is the ability to calculate the incremental likelihood $p\left(y_t|\mathcal{D}_{t-1}\right)$. This incremental likelihood, specified in (3.7) as

$$p\left(y_t|\mathcal{D}_{t-1}\right) = \int p\left(y_t|\boldsymbol{\theta}_t\right) p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\right) p\left(\boldsymbol{\theta}_{0:t-1}|\mathcal{D}_{t-1}\right) d\boldsymbol{\theta}_{0:t}, \tag{15.2}$$

will not be tractable for most of our models of interest (with the exception of the NDLM). As discussed in Section 6.1, particle filters give an unbiased estimator of the marginal likelihood, such that (given that $\left\{w_t^{(i)}\right\}_{i=1}^{N_p}$ are unnormalised importance weights)

$$\hat{\ell}_t\left(\boldsymbol{\Phi}\right) = \hat{p}\left(y_t|\mathcal{D}_{t-1}, \boldsymbol{\Phi}\right) = \frac{1}{N_p}\sum_{i=1}^{N_p} w_t^{(i)}.$$

Consequently, as

$$p\left(\mathcal{D}_t|\boldsymbol{\theta}_{0:t}\right) = p\left(y_1\right)\prod_{k=2}^{t} p\left(y_k|\mathcal{D}_{k-1}\right),$$

we can then write

$$\hat{\ell}_{0:t}\left(\boldsymbol{\Phi}\right) = \hat{p}\left(\mathcal{D}_t|\boldsymbol{\theta}_{0:t}, \boldsymbol{\Phi}\right) = \prod_{k=1}^{t}\left(\frac{1}{N_p}\sum_{n=1}^{N_p} w_k^{(i)}\right). \tag{15.3}$$

Using this property, SMC^2 removes the IBIS limitation of the incremental likelihood tract-

ability by coupling IBIS with particle filtering methods as an unbiased estimator of the incremental likelihood in (15.2) making it is possible to adapt IBIS to a wide class of models, specifically of interest in the context of this thesis, to non-linear DGLMs.

Whereas IBIS makes use of the exact solution to calculate the incremental likelihoods $\left\{p\left(y_t|y_{t-1}, \boldsymbol{\Phi}^{(m)}\right)\right\}_{m=1}^{N_{\boldsymbol{\Phi}}}$ by using $N_{\boldsymbol{\Phi}}$ Kalman filters, SMC^2 will then replace this exact calculation with an approximation provided by the unbiased estimate from a particle filter. Each of the $N_{\boldsymbol{\Phi}}$ particles will have an associated particle filter, itself consisting of $N_{\boldsymbol{\theta}}$ particles, enabling the estimation of the likelihood in (15.3) according to Algorithm 15.1.

Using the same notation as in Chopin *et al.* (2013), we consider a generic implementation of an SIR filter as described in Algorithm 15.1. To clarify the distinction between the state particle importance weights and the parameter particles weights we will refer to them respectively as $w_t^{\boldsymbol{\theta}(i)}$, for $i = 1, \ldots, N_{\boldsymbol{\theta}}$, and $w_t^{\boldsymbol{\Phi}(m)}$, for $m = 1, \ldots, N_{\boldsymbol{\Phi}}$. All the steps of the m^{th}, $m = 1, \ldots, N_{\boldsymbol{\Phi}}$, particle filter are conditioned on a particular value of the parameter set $\boldsymbol{\Phi}^{(m)}$ and we will also adopt Chopin *et al.* (2013) notation for the resampled indices where $a_{t-1}^{(i)}$ denotes the ancestor of particle i at time t. Regarding the resampling method, although Chopin *et al.* (2013) considers Multinomial[1] resampling, any unbiased resampling method, such as the ones described in Section 7.3 on page 87 and generically referred to as $\mathcal{R}(\cdot)$, can be implemented.

Within the context of SMC^2 the particle filter serves two main purposes. Firstly, to estimate the incremental likelihood in (15.3), up until time t and conditioned on a fixed $\boldsymbol{\Phi}^{(m)}$. Secondly, it will return the new state particles along with their respective ancestors, that is $\left\{a_{1:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}$.

According to Chopin *et al.* (2013), each of the $\left\{\boldsymbol{\Phi}^{(m)}\right\}_{m=1}^{N_{\boldsymbol{\Phi}}}$ particles will explore the parameter space and at each time point t will be weighted according to its respective m^{th} particle filter's likelihood estimation, $\hat{\ell}_t\left(\boldsymbol{\Phi}^{(m)}\right)$ and resampled if a degeneracy criterion is met.

For SMC^2 the same degeneracy criterion as IBIS (described in (14.1)) is used. If it is below a pre-defined threshold, ESS_{eff}, we resample the $\boldsymbol{\Phi}$-particles using an unbiased resampling method as mentioned previously.

Considering that at time t we are in possession of the particle filter's result for each of the $m = 1, \ldots, N_{\boldsymbol{\Phi}}$ $\boldsymbol{\Phi}$-particles, that is $\left\{\left[\left\{a_{1:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}, w_{0:t}^{\boldsymbol{\theta}(i)}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}\right]^{(m)}\right\}_{m=1}^{N_{\boldsymbol{\Phi}}}$, we can then resample the $\boldsymbol{\Phi}$-particles by using their weights, calculated recursively as

[1]Described in section 7.3.1 on page 89.

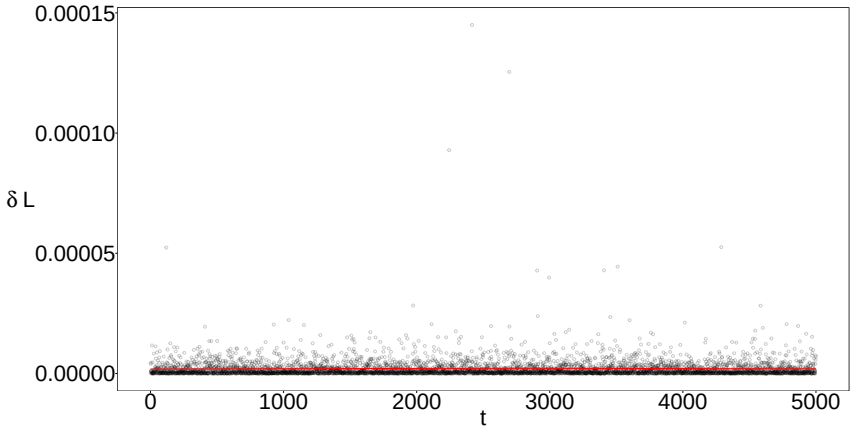


Figure 15.1: $\left(p\left(y_t|y_{t-1},\boldsymbol{\Phi}\right)-\hat{p}\left(y_t|y_{t-1},\boldsymbol{\Phi}\right)\right)^2$ for a $\mathcal{M}=\{\mathcal{P}(1)\}$ NDLM using a Kalman filter and a SIR filter filter for the incremental likelihood estimate.

Algorithm 15.1 Generic *sample-resample* PF

initialisation

$$\left\{\boldsymbol{\theta}_0^{(i)}\right\}_{i=1}^{N_\theta} \sim p\left(\boldsymbol{\theta}\right)$$
$$\left\{a_0^{(i)}\right\}_{i=1}^{N_\theta} = i$$

for $t \leftarrow 1$ to k

sample $\boldsymbol{\theta}_t^{(i)} \sim \pi\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{a_{t-1}^{(i)}}, y_t, \boldsymbol{\Phi}\right)$

calculate weights

$$w_t^{\boldsymbol{\theta}(i)} = w_{t-1}^{\boldsymbol{\theta}(i)} \frac{p\left(y_t|\boldsymbol{\theta}_t^{(i)},\boldsymbol{\Phi}\right) p\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{a_{t-1}^{(i)}},\boldsymbol{\Phi}\right)}{\pi\left(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{a_{t-1}^{(i)}}, y_t, \boldsymbol{\Phi}\right)}$$

normalise weights

$$\tilde{w}_t^{\boldsymbol{\theta}(i)} = \frac{w_t^{\boldsymbol{\theta}(i)}}{\sum_{i=1}^{N_{\boldsymbol{\theta}}} w_t^{\boldsymbol{\theta}(i)}}$$

sample ancestors

$$a_t^{(i)} \sim \mathcal{R}\left(\left\{w_t^{\boldsymbol{\theta}(i)}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}\right)$$

calculate incremental likelihood

$$\hat{\ell}_t\left(\boldsymbol{\Phi}\right) = \frac{1}{N_{\boldsymbol{\theta}}} \sum_{i=1}^{N_{\boldsymbol{\theta}}} w_t^{\boldsymbol{\theta}(i)}$$

$$w_t^{\boldsymbol{\Phi}(m)} = w_{t-1}^{\boldsymbol{\Phi}(m)} \hat{\ell}_t \left(\boldsymbol{\Phi}^{(m)} \right).$$

The resulting resampled $\boldsymbol{\Phi}$-particles, $\boldsymbol{\Phi}^{(m)^{(n)}}$ will then be propagated from $p\left(\boldsymbol{\Phi}|\mathcal{D}_{t-1}\right)$ to $p\left(\boldsymbol{\Phi}|\mathcal{D}_t\right)$ using a PMCMC step as described in Section 13 on page 152. Although in Chopin *et al.* (2013) a PMMH step is used, presented generically in Algorithm 15.2, other PMCMC kernels may be used if possible (such as Particle Gibbs). The PMMH proposal could consist, for instance, of a random-walk proposal, such that

$$\begin{aligned} \boldsymbol{\Phi}^{\star(m)} &\sim \mathcal{N}\left(\boldsymbol{\Phi}^{(m)^{(n)}}, \Sigma \right) \\ \Sigma &= \text{Var}\left[\boldsymbol{\Phi}^{(m)^{(n)}} | \mathcal{D}_t \right]. \end{aligned}$$

Algorithm 15.2 Particle Markov Metropolis-Hastings

for $i \leftarrow 1$ to $N_{\boldsymbol{\Phi}}$

sample $\boldsymbol{\Phi}^{\star(i)} \sim p\left(\boldsymbol{\Phi}^{\star}|\boldsymbol{\Phi}^{(i)}\right)$

calculate $\hat{\ell}_{0:t}\left(\boldsymbol{\Phi}^{\star(i)}\right)$ using a particle filter as in Algorithm 15.1

calculate acceptance rate

$$A^{(i)} = \min\left\{ 1, \frac{\hat{\ell}_{0:t}\left(\boldsymbol{\Phi}^{\star(i)}\right) p\left(\boldsymbol{\Phi}^{\star(i)}\right)}{\hat{\ell}_{0:t}\left(\boldsymbol{\Phi}^{(i)}\right) p\left(\boldsymbol{\Phi}^{(i)}\right)} \right\}$$

sample $u \sim \mathcal{U}(0,1)$

if $u < A^{(i)}$

  set $\boldsymbol{\Phi}^{(i)} = \boldsymbol{\Phi}^{\star(i)}$

else

  set $\boldsymbol{\Phi}^{(i)} = \boldsymbol{\Phi}^{(i)}$

Chopin *et al.* (2013) notes that if we extend our definition of the likelihood approximation in (15.3) as function of the particle filter's returned quantities as

$$\begin{aligned} \hat{\ell}_{0:t}\left(\boldsymbol{\Phi}, \left\{ \boldsymbol{\theta}_{0:t}^{(i)}, a_{1:t}^{(i)} \right\}_{i=1}^{N_{\boldsymbol{\theta}}}\right) &= \prod_{k=0}^{t} \hat{\ell}_k\left(\boldsymbol{\Phi}\right) \\ &= \left[\frac{1}{N_{\boldsymbol{\theta}}} \sum_{i=1}^{N_{\boldsymbol{\theta}}} w_0^{\boldsymbol{\Phi}}\left(\boldsymbol{\theta}_0^{(i)}\right) \right] \prod_{k=1}^{t} \left\{ \frac{1}{N_{\boldsymbol{\theta}}} \sum_{n=1}^{N_{\boldsymbol{\theta}}} w_k^{\boldsymbol{\Phi}}\left(\boldsymbol{\theta}_{k-1}^{(a_k^i)}, \boldsymbol{\theta}_k^{(i)}\right) \right\} \end{aligned} \quad (15.4)$$

the PMMH acceptance ratio is calculated according to

$$\min\left\{1, \frac{p\left(\boldsymbol{\Phi}^{\star}\right)\hat{\ell}_{0:t}\left(\boldsymbol{\Phi}^{\star}, \left\{\boldsymbol{\theta}_{0:t}^{\star(i)}, a_{1:t}^{\star(i)}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}\right)}{p\left(\boldsymbol{\Phi}\right)\hat{\ell}_{0:t}\left(\boldsymbol{\Phi}, \left\{\boldsymbol{\theta}_{0:t}^{(i)}, a_{1:t}^{(i)}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}\right)}\right\}. \tag{15.5}$$

Since the likelihood estimate for each of the m^{th} current particle, $\hat{\ell}_{0:t}\left(\boldsymbol{\Phi}, \left\{\boldsymbol{\theta}_{0:t}^{(i)}, a_{1:t}^{(i)}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}\right)$, can be calculated sequentially as in (15.3), SMC² requires only the storage of each m $\left\{\boldsymbol{\theta}_{t-1}^{(i)}, a_{t-1}^{(i)}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}$ and $\boldsymbol{\Phi}^{(m)}$, the algorithm (as IBIS) can be considered sequential and online when no $\boldsymbol{\Phi}$-particle resampling and PMCMC is occurring and sequential *(i.e.* running at $\mathcal{O}\left(tN_{\Phi}N_{\theta}\right)$), but not online, when resampling occurs and the PMCMC step takes place (running at $\mathcal{O}\left(N_{\Phi}N_{\theta}\right)$). Moreover, as with IBIS, if we were interested solely in parameter estimation, the storage of the running total of the marginal likelihoods would suffice to calculate the PMMH acceptance ratio. However, as mentioned previously, we are are concerned with the estimation of both state and parameters in this thesis.

As seen previously, SMC² is a versatile method since few assumptions are made for many of the components, making them implementation agnostic. For instance, although a generic implementation of a SIR particle filter is given in Algorithm 15.1, any implementation of a particle filter, as long as it provides an unbiased estimate of $p\left(\boldsymbol{\theta}_0|\mathcal{D}_t, \boldsymbol{\Phi}\right)$ and $\hat{\ell}_{0:t}\left(\boldsymbol{\Phi}\right)$ can be used. This includes all of the methods described in Section 7.4, namely SIR and the APF. The resampling method $\mathcal{R}\left(\cdot\right)$, in Algorithm 15.1, is also left to our choice, provided it performs an unbiased selection of particles based on the weights, giving the possibility to implement any of the methods described in Section 7.3. The Particle Markov Chain Monte Carlo (PMCMC) kernel can also be implemented as a Metropolis-Hasting or, for instance, as previously mentioned, a Particle Gibbs kernel (Chopin *et al.* (2015)). Furthermore, for the PMMH, we could also use a random-walk proposal, such as in Algorithm 15.2 or an independent kernel.

If, as in Chopin *et al.* (2013), we denote the joint probability of the random variables returned by the particle filter, conditioned on a certain $\boldsymbol{\Phi}$, that is $\left\{a_{1:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}\Big|\boldsymbol{\Phi}$, as

$$\psi_t^{\boldsymbol{\Phi}}\left(\left\{a_{1:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}\right),$$

SMC² will then target[2]

$$\pi_t\left(\boldsymbol{\Phi}, \left\{a_{1:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right\}_{i=1}^{N_{\theta}}\right) = \frac{p\left(\boldsymbol{\Phi}\right)}{p\left(\mathcal{D}_t\right)}\psi_t^{\boldsymbol{\Phi}}\left(\left\{a_{1:t}^{(i)}, \boldsymbol{\theta}_{0:t}^{(i)}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}\right)\prod_{k=0}^{t}\hat{\ell}_k\left(\boldsymbol{\Phi}\right), \tag{15.6}$$

[2]For a formal justification *cf.* Chopin *et al.* (2013).

where π_t has a marginal distribution with respect to $\boldsymbol{\Phi}$ of $p(\boldsymbol{\Phi}|\mathcal{D}_t)$ (Chopin *et al.* (2013)).

As noted in Chopin *et al.* (2013), the choice of an appropriate $N_{\boldsymbol{\theta}}$ and $N_{\boldsymbol{\Phi}}$ is not trivial and that to maintain reasonable acceptance rates in the rejuvenation step we should aim at $N_{\boldsymbol{\theta}} = \mathcal{O}(t)$ (Andrieu *et al.* (2010)). A solution proposed in Chopin *et al.* (2013) is to start with a low $N_{\boldsymbol{\theta}}$ and increase it by a determined factor[3] whenever the MH acceptance falls beneath a predefined threshold. Although this method can improve the MH acceptance rate, this was not incorporated in the SMC² (or O-SMC², in the following section) implementation used in the thesis.

A general algorithm for SMC² is presented in Algorithm 15.3.

Example. SMC² for an AR(1) Poisson DLM.

The Autoregressive of order 1 (AR(1)) Poisson DLM can be specified as

$$\begin{aligned} y_t|\theta_t &\sim \text{Po}(\lambda_t) \\ \lambda_t &= \exp\{\theta_t\} \\ \theta_t|\theta_{t-1}, \Phi &\sim \mathcal{N}\left(\alpha + \beta\theta_{t-1}, \tau^2\right) \end{aligned}$$

where α and β are regression coefficients, in this case considered static but unknown. The parameter set used for the realisation of the $N_{obs} = 2000$ observations in Figure 15.2a was $\boldsymbol{\Phi} = \{\tau^2, \alpha, \beta\} = \{1.5, 0.7, 0.3\}$. For this model we can implement SMC² using a standard bootstrap filter (*i.e.* a SIR filter using the model's transition $p(\theta|\theta_{t-1})$ as the importance density) to approximate the likelihood. The priors used were

$$\begin{aligned} \tau_0^2 &\sim \mathcal{IG}(1,1) \\ \alpha_0 &\sim \text{Beta}(1,1) \\ \beta_0 &\sim \text{Beta}(1,1). \end{aligned}$$

The SMC² estimation was performed using $N_{\boldsymbol{\Phi}} = 1000$ and $N_{\boldsymbol{\theta}} = 2000$ particles. We can see the estimation history for τ^2, α and β respectively in Figures 15.3a, 15.3b and 15.3c with the horizontal dashed line representing the true values. Also in Figures 15.3a, 15.3b and 15.3c we can see the marginal for τ^2, α and β at the final time point, $t = 2000$, with the vertical dashed lines representing the true values for comparison.

[3] In Chopin *et al.* (2013) the example of increasing $N_{\boldsymbol{\theta}}$ by a factor of 2 is given.

Algorithm 15.3 SMC²

sample

$$\left\{\boldsymbol{\Phi}_0^{(m)}\right\}_{m=1}^{N_{\Phi}} \sim p\left(\boldsymbol{\Phi}\right)$$
$$\{w_0\}_{i=1}^{N_{\Phi}} = 1$$

for $t \leftarrow 1$ to N_{obs}

for $m \leftarrow 1$ to $N_{\boldsymbol{\Phi}}$

calculate using a particle filter (as in Algorithm 15.1)

$$\hat{p}\left(y_t|\mathcal{D}_{t-1}, \boldsymbol{\Phi}_{t-1}^{(m)}\right) = \frac{1}{N_{\boldsymbol{\theta}}} \sum_{i=1}^{N_{\boldsymbol{\theta}}} w_t^{\boldsymbol{\theta}(i)}$$

update the $\boldsymbol{\Phi}$ importance weights as

$$w_t^{\boldsymbol{\Phi}(m)} = w_{t-1}^{\boldsymbol{\Phi}(m)} \hat{p}\left(y_t|\mathcal{D}_{t-1}, \boldsymbol{\Phi}_{t-1}^{(m)}\right)$$

calculate the ESS as

$$\widehat{ESS}_t = \frac{\left(\sum_{m=1}^{N_{\boldsymbol{\Phi}}} w_t^{\boldsymbol{\Phi}(m)}\right)^2}{\sum_{m=1}^{N_{\boldsymbol{\Phi}}} \left(w_t^{\boldsymbol{\Phi}(m)}\right)^2}$$

if $\widehat{ESS}_t < ESS_{eff}$

resample $\boldsymbol{\Phi}$ particles according to

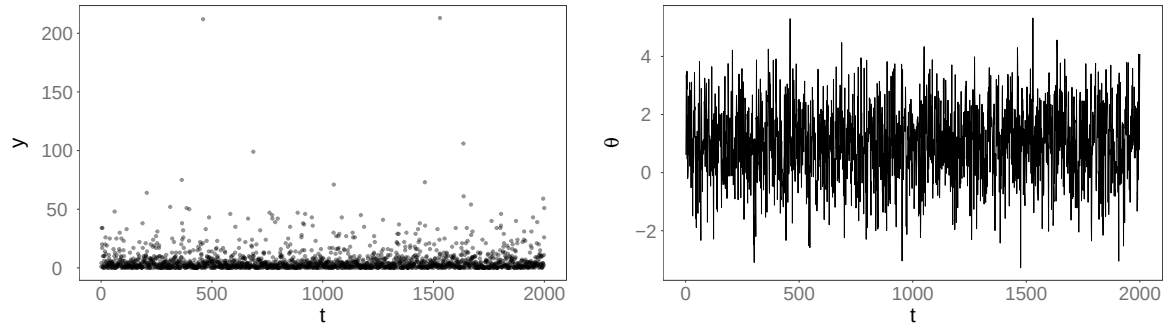
$$k \sim \mathcal{R}\left(\left\{w_t^{\boldsymbol{\Phi}(m)}\right\}_{m=1}^{N_{\boldsymbol{\Phi}}}\right)$$
$$\left\{\boldsymbol{\Phi}_{t-1}^{(m)}, \left\{\boldsymbol{\theta}_{0:t}^{(i)}, a_{1:t}^{(i)}\right\}^{(m)}, w_t^{\boldsymbol{\Phi}(m)}\right\}_{m=1}^{N_{\boldsymbol{\Phi}}} \leftarrow \left\{\boldsymbol{\Phi}_{t-1}^{(k)}, \left\{\boldsymbol{\theta}_{0:t}^{(i)}, a_{1:t}^{(i)}\right\}^{(k)}, 1\right\}_{m=1}^{N_{\boldsymbol{\Phi}}}$$

run one step of a PMCMC (as in Algorithm 15.2) to sample

$$\left\{\boldsymbol{\Phi}_t^{(m)}\right\}_{m=1}^{N_{\boldsymbol{\Phi}}} \sim p\left(\boldsymbol{\Phi}_t|\boldsymbol{\Phi}_{t-1}^{(m)}, \mathcal{D}_t\right)$$

else

set $\left\{\boldsymbol{\Phi}_t^{(m)}\right\}_{m=1}^{N_{\boldsymbol{\Phi}}} \leftarrow \left\{\boldsymbol{\Phi}_{t-1}^{(m)}\right\}_{m=1}^{N_{\boldsymbol{\Phi}}}$



(a) $\mathcal{D}_t$ (b) $\theta_{0:t}$

Figure 15.2: Observations (*left*) and states (*right*) from a realisation of AR(1) PoDLM with $\boldsymbol{\Phi} = \{\tau^2, \alpha, \beta\} = \{1.5, 0.7, 0.3\}$.

15.1 Online SMC^2

As mentioned previously, SMC^2 is not a true online filter. The necessity of step (15.4) (where as soon as $\widehat{ESS}_{\boldsymbol{\Phi}}$ falls beneath a certain threshold, a PMCMC step will be triggered) involving a calculation of total likelihood means that the computational costs will grow with t, although, theoretically (Chopin *et al.* (2013)) this step occurs with decreasing frequency.

An *ad-hoc* solution to implement SMC^2 in an online manner is to restrict the likelihood calculation used in PMMH set to a fixed size subset of past data, keeping the computational and storage costs bounded (*i.e.* changing the cost from[4] $\mathcal{O}(t)$ to $\mathcal{O}(1)$ for each iteration). By considering a fixed window of observations of $k > 0$, such that $\tilde{Y}_t^k = \{y_{t-k}, y_{t-k+1}, \ldots, y_t\}$ we can see that the online part of SMC^2 will not be altered, however, when the rejuvenation process happens we will now be targeting the parameter posterior

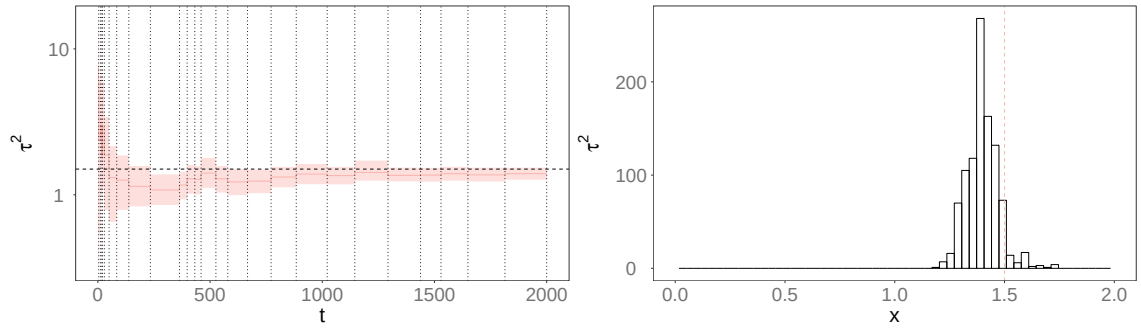
$$p\left(\boldsymbol{\Phi}|\tilde{Y}_t^k\right).$$

This online approximation will target the correct posterior of the parameters within the observation window $\tilde{Y}_t^k$. That is, by definition, this method will target the posterior of the parameter set $\boldsymbol{\Phi}$, conditioned on the observations $\tilde{Y}_t^k$. However, we will assume that

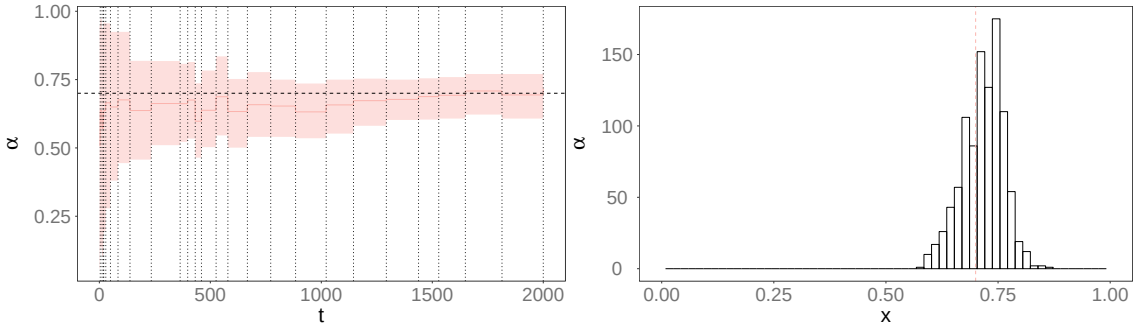
$$p\left(\boldsymbol{\Phi}|\tilde{Y}_t^k\right) \approx p\left(\boldsymbol{\Phi}|\mathcal{D}_t\right).$$

With this approximation we will lose the property in (15.1), that is, the ability to estimate sequentially the distributions $\{p(\boldsymbol{\Phi}), p(\boldsymbol{\Phi}|y_1), p(\boldsymbol{\Phi}|y_{1:2}), \ldots, p(\boldsymbol{\Phi}|\mathcal{D}_t)\}$, however,

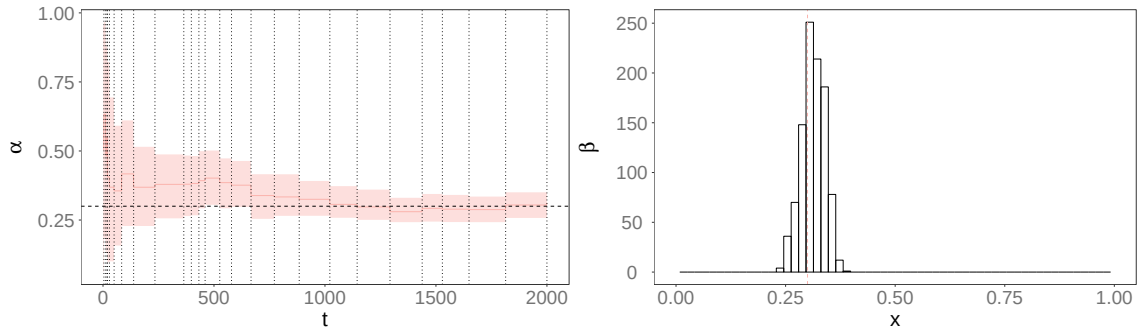
[4]To be precise, for a model with fixed $N_{\boldsymbol{\Phi}}$ and $N_{\boldsymbol{\theta}}$, the cost at each iteration would be $\mathcal{O}(tN_{\boldsymbol{\Phi}}N_{\boldsymbol{\theta}})$ and this would become (with an observation window k) at most $\mathcal{O}(kN_{\boldsymbol{\Phi}}N_{\boldsymbol{\theta}})$, where $k, N_{\boldsymbol{\Phi}}$ and $N_{\boldsymbol{\theta}}$ are constant.



(a) τ^2 estimation history (*left*) and marginals at time $t = 2000$ (*right*)



(b) α estimation history (*left*) and marginal at $t = 2000$ (*right*)



(c) β estimation history (*left*) and marginal at $t = 2000$ (*right*)

Figure 15.3: Posterior estimation history (*left column*, colour lines represent the posterior mean and shaded area the 90% equitailed credibility interval) of the parameters $\boldsymbol{\Phi} = \left\{\tau^2, \alpha, \beta\right\}$ and posteriors at time $t = N_{obs}$ (*right column*, vertical red line represents the "true" value) of a AR(1) PoDLM using SMC^2 with $N_{\boldsymbol{\Phi}} = 1000$ and $N_{\boldsymbol{\theta}} = 2000$.

for the purposes of online inference with special focus on state and observations forecast as well as anomaly detection, an approximation of $p(\boldsymbol{\Phi}|\mathcal{D}_t)$ will suffice, especially if we are skeptical about the assumption that the "static" parameters will remain constant indefinitely.

Regarding the PMCMC step (and once we've reached the $t > k$ threshold), the MH acceptance will not be in the form of (15.5) but instead

$$\min\left\{1, \frac{p(\boldsymbol{\Phi}^\star)\,\hat{\ell}_{t-k:t}\left(\boldsymbol{\Phi}^\star, \left\{\boldsymbol{\theta}^{\star(i)}_{t-k:t}, a^{\star(i)}_{t-k+1:t}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}\right)}{p(\boldsymbol{\Phi})\,\hat{\ell}_{t-k:t}\left(\boldsymbol{\Phi}, \left\{\boldsymbol{\theta}^{(i)}_{t-k:t}, a^{(i)}_{t-k:t}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}\right)}\right\}. \tag{15.7}$$

Whereas in the "full" SMC², in the sequential phase, the total likelihood could be calculated recursively as in (15.4), that is

$$\hat{\ell}_{0:t}\left(\boldsymbol{\Phi}, \left\{\boldsymbol{\theta}^{(i)}_{0:t}, a^{(i)}_{1:t}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}\right) = \hat{\ell}_0(\boldsymbol{\Phi}) \times \hat{\ell}_1(\boldsymbol{\Phi}) \times \cdots \times \hat{\ell}_t(\boldsymbol{\Phi})$$

this will now not be possible, since we need to remove (assuming $t > k$) the terms $\hat{\ell}_0(\boldsymbol{\Phi}), \ldots, \hat{\ell}_{t-k-1}(\boldsymbol{\Phi})$ for the partial likelihood calculation in (15.7). This can be achieved by simply keeping a First In, First Out (FIFO) queue of maximum size k such that

$$Q = \left[\hat{\ell}_{t-k}(\boldsymbol{\Phi}), \hat{\ell}_{t-k+1}(\boldsymbol{\Phi}), \ldots, \hat{\ell}_t(\boldsymbol{\Phi})\right], \qquad t > k$$

and generating the partial likelihood as

$$\hat{\ell}_{t-k:t}(\boldsymbol{\Phi}) = \prod_{n=1}^{k} Q_n$$

when needed for the calculation of the MH acceptance ratio such that

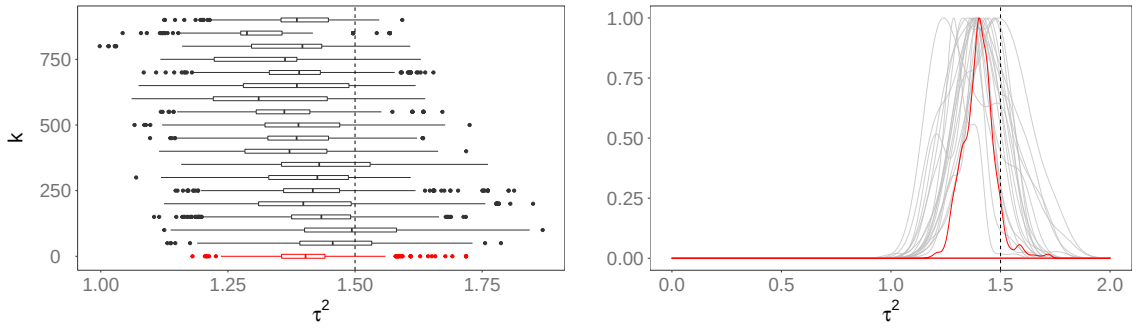
$$\min\left\{1, \frac{p(\boldsymbol{\Phi}^\star)\,\hat{\ell}_{t-k:t}\left(\boldsymbol{\Phi}^\star, \left\{\boldsymbol{\theta}^{\star(i)}_{t-k:t}, a^{\star(i)}_{t-k+1:t}\right\}_{i=1}^{N_{\boldsymbol{\theta}}}\right)}{p(\boldsymbol{\Phi})\prod_{n=1}^{k} Q_n}\right\}.$$

This approximation was applied to the model and data for the AR(1) PoDLM on page 182 using the same priors and particle sizes and with a range of window sizes $k = 50, 100, 150, \ldots, 500$. We can see from Figure 15.4 that for all parameters $\boldsymbol{\Phi} = \{\tau^2, \alpha, \beta\}$ there is a considerable overlap of the $\boldsymbol{\Phi}$-particles, even when using very small window sizes ($k = 50$). The mean value of the estimated parameters is generally consistent with the full SMC² estimation (in red) and the most notable difference is an increased variance in the

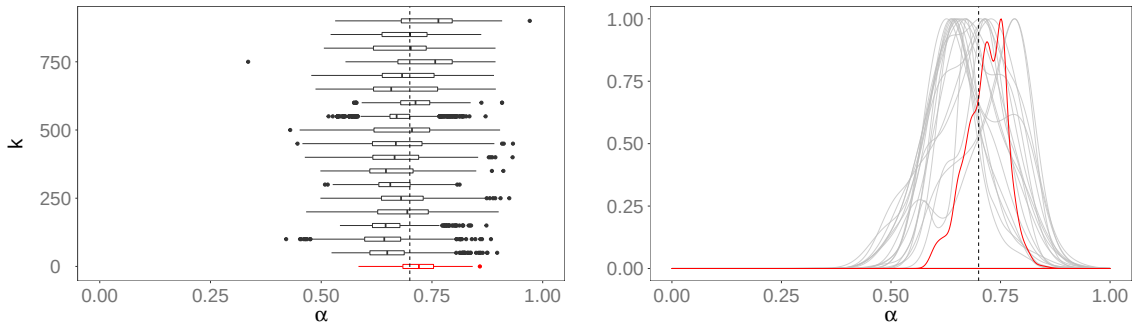
k	$\overline{\tau}^2$	$\overline{\alpha}$	$\overline{\beta}$	k	$\overline{\tau}^2$	$\overline{\alpha}$	$\overline{\beta}$
50	1.467 (0.1162)	0.6522 (0.0619)	0.294 (0.0395)	500	1.4011 (0.1154)	0.6832 (0.0872)	0.3129 (0.0594)
100	1.4946 (0.1331)	0.6403 (0.069)	0.3117 (0.0351)	550	1.3583 (0.0862)	0.6725 (0.0588)	0.3224 (0.036)
150	1.4311 (0.0962)	0.6507 (0.0497)	0.3387 (0.035)	600	1.3317 (0.1281)	0.7093 (0.0568)	0.3062 (0.0435)
200	1.4164 (0.1404)	0.6826 (0.084)	0.3081 (0.0566)	650	1.379 (0.1232)	0.6846 (0.0861)	0.3105 (0.0491)
250	1.415 (0.0994)	0.6862 (0.0731)	0.3241 (0.0507)	700	1.384 (0.0826)	0.693 (0.0694)	0.3053 (0.0389)
300	1.408 (0.0981)	0.6661 (0.0521)	0.3267 (0.0361)	750	1.3234 (0.0961)	0.7308 (0.0715)	0.3095 (0.0512)
350	1.446 (0.1231)	0.6577 (0.065)	0.3116 (0.0578)	800	1.3856 (0.1134)	0.6842 (0.0897)	0.3575 (0.0444)
400	1.3683 (0.1091)	0.6681 (0.0768)	0.3453 (0.0475)	850	1.3091 (0.0693)	0.6806 (0.0662)	0.307 (0.0378)
450	1.3863 (0.0892)	0.6724 (0.0799)	0.311 (0.0449)	900	1.3891 (0.0763)	0.7297 (0.0841)	0.3021 (0.0477)
SMC²	1.4006 (0.0725)	0.7168 (0.0454)	0.3126 (0.0261)	SMC²	1.4006 (0.0725)	0.7168 (0.0454)	0.3126 (0.0261)

Table 15.1: Summary of parameter posterior mean and standard deviation (in brackets) for τ^2, α and β using O-SMC^2 with different window sizes k. SMC^2 posterior mean and standard deviation included for comparison.

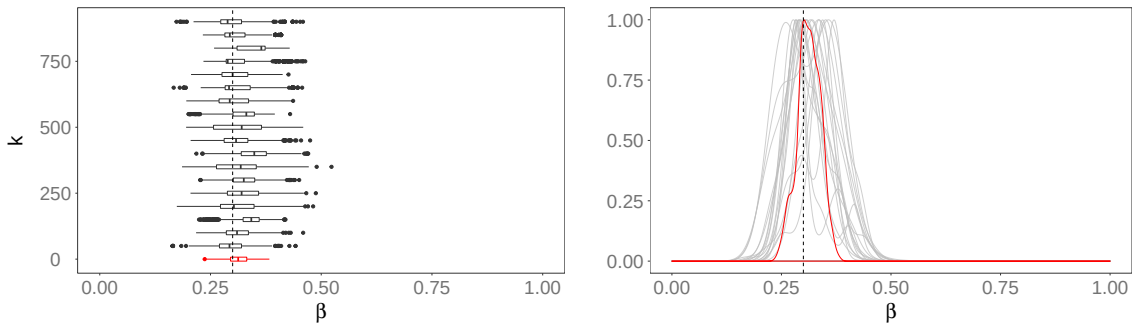
parameter estimation.



(a) τ^2 marginals at $t = 2000$ for different k window sizes



(b) α^2 marginals at $t = 2000$ for different k window sizes



(c) β^2 marginals at $t = 2000$ for different k window sizes

Figure 15.4: τ^2, α and β posteriors at $t = N_{obs}$ for a AR (1) PoDLM with O-SMC^2 using different k window sizes. Dashed lines represent truth. Full SMC^2 estimation in red.

Part V

Case Studies

Chapter 16

Results

16.1 Particle numbers

One of issues with SMC estimation methods is the one of Monte Carlo errors. Due to the accumulations of error inherent to the discrete posterior approximation using a finite numbers of particles, variability will inevitably occur between runs for the same model even considering the same priors and initial conditions. This problem is even more evident when dealing with a large number of observations (which could even lead to the collapse of the filters).

To determine the importance of these errors in the final estimation values and to provide a starting point to determine an adequate number of particles for the subsequent analyses in this chapter, we consider the case of online parameter estimation and to quantify the variability, the mean Monte Carlo Mean Absolute Error (MCMAE) was used. If we consider $\overline{\boldsymbol{\Phi}}$ as the parameter estimate from the PMMH runs as the reference, we calculate the estimation history's absolute error for a particle filter as

$$MCMAE_{\boldsymbol{\Phi}} = \frac{1}{N_{obs}} \sum_{t=1}^{N_{obs}} \left| \overline{\boldsymbol{\Phi}} - \widehat{\boldsymbol{\Phi}}_t \right|, \tag{16.1}$$

where $\widehat{\boldsymbol{\Phi}}_t$ is the filter's estimation of the parameter posterior mean at time t and $\overline{\boldsymbol{\Phi}}$ is the PMMH posterior mean estimation. The mean $MCMAE$ will then be the averaged value for n runs, such that

$$\overline{MCMAE}_{\boldsymbol{\Phi}} = \frac{1}{n} \sum_{i=1}^{n} MCMAE_{\boldsymbol{\Phi},i}. \tag{16.2}$$

Regarding the variability of the state estimation, the mean $MCMAE$ can also be calculated. In this case, the PMMH smoothed state estimate is used as the reference, such

that

$$MCMAE_{\boldsymbol{\theta}} = \frac{1}{N_{obs}} \sum_{t=1}^{N_{obs}} \left| \overline{\boldsymbol{\theta}}_t - \widehat{\boldsymbol{\theta}}_t \right|, \tag{16.3}$$

where $\widehat{\boldsymbol{\theta}}_t$ is the filter's estimation of the state posterior mean at time t and $\overline{\boldsymbol{\theta}}$ is the PMMH posterior mean estimation. As in the parameter case, the mean $MCMAE$ will then be the averaged value for n runs.

To illustrate the effect of the number of particles in the performance of the discussed SMC methods, we perform state and parameter estimation on a realisation of $N_{obs} = 1000$ observations of a $\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(500, 1)\}$ PoDLM. The parameter set is $\boldsymbol{\Phi} = \{\mathsf{W}\}$ with

$$\mathsf{W} = \text{diag}(0.1, 0.1, 0.1)$$

and state priors

$$\boldsymbol{\theta}_0 \sim \mathcal{N}\left([3 \quad 0 \quad 0]^T, \text{diag}(10, 10, 10) \right),$$

and parameter priors

$$p(\mathsf{W}_0) = \mathcal{IW}(3, \mathbf{I}_3).$$

The filters used were LW (as described in Chapter 8 on page 106), Storvik (Section 9 on page 116) and PL (Section 10 on page 125). Storvik was implemented used three different importance densities, respectively, sampling from the prior (Storvik, Section 6.2.2 on page 76), a CF proposal (Storvik-CF, Section 6.2.4 on page 79) and a EKF proposal (Storvik-EKF, Section 6.2.5 on page 80). Resampling was performed using Stratified resampling with a static checkpoint $n = 1$. By using a static checkpoint we can compare iteration times (the computational cost of each online filter at each time step t) between filters without considerations of how frequently the resampling stage occurs. The state $\overline{MCMAE}_{\boldsymbol{\theta}}$ (relatively to a PMMH long run[1]) was calculated for these methods using different particle numbers, ranging from $100 \leq N_p \leq 4 \times 10^4$ and the results for each state component can be seen in Figures 16.1, 16.2 and 16.3. It is clear from the results that initially, with a number of particles as low as $N_p = 100$ there is a very high variability between state estimates when compared to the "true" values, independently of the method used. As expected, the variability decreases with the increase in particle numbers appearing to stabilise, for this particular model, around $N_p = 3 \times 10^4$. We decided however to use a lower value of $N_p = 2 \times 10^4$ for the subsequent analyses with the intent to better highlight potential differences between the different SMC methods while still keeping a considerable amount of particles to avoid filter collapse.

[1] Trace and ACF plots can be viewed in Appendix A.3.

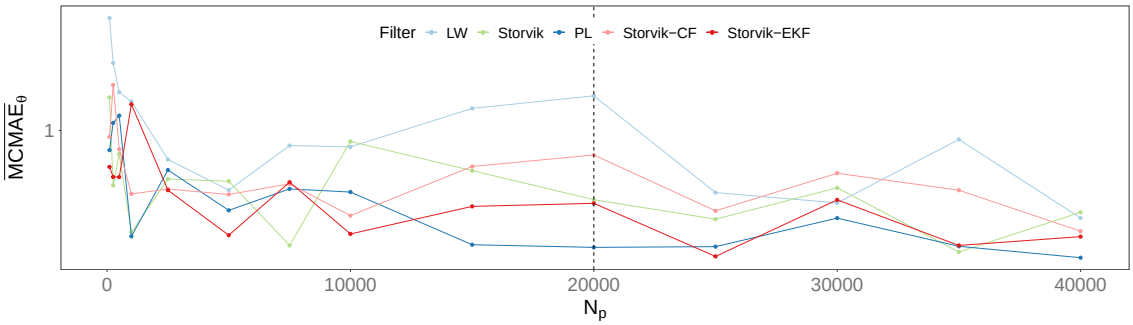


Figure 16.1: $\overline{MCMAE}\,(\theta_1, k)$ for $100 \leq k \leq 4 \times 10^4$ with $n = 10$ runs (*log scale*)

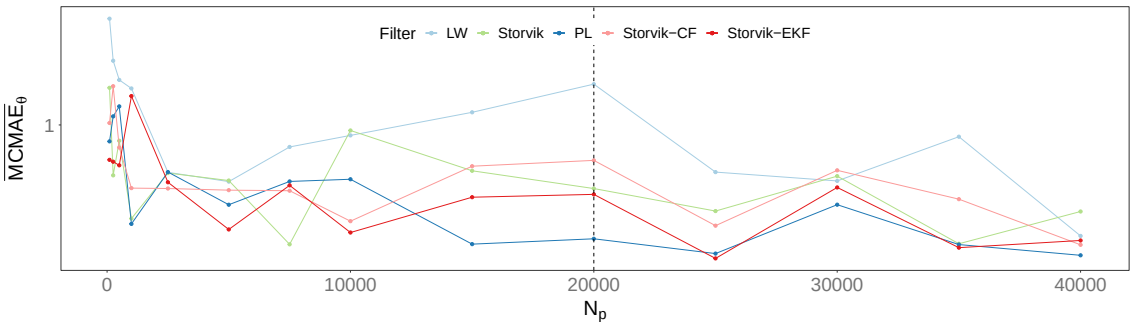


Figure 16.2: $\overline{MCMAE}\,(\theta_2, k)$ for $100 \leq k \leq 4 \times 10^4$ with $n = 10$ runs (*log scale*)

16.2 Resampling algorithms

To benchmark the resampling methods presented in Section 7.3 on page 87, a bootstrap filter was used in a realisation of $N_{obs} = 500$ observations of a AR(1) Poisson DLM (presented in Figure 16.4) with $\boldsymbol{\Phi} = \left\{\alpha, \beta, \tau^2\right\} = \{0.5, 0.5, 1\}$, that is:

$$
\begin{aligned}
y_t &\sim \text{Po}\left(e^{\theta_t}\right) \\
\theta_t &\sim \mathcal{N}\left(0.5 + 0.5\theta_{t-1}, 1\right).
\end{aligned}
$$

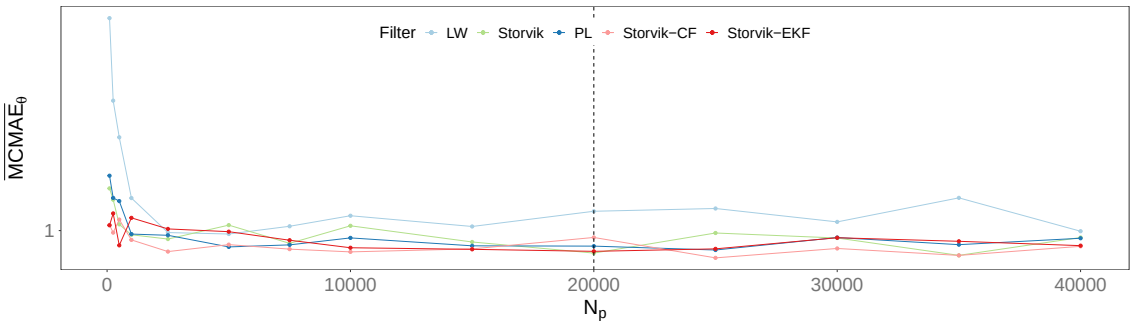


Figure 16.3: $\overline{MCMAE}\,(\theta_3, k)$ for $100 \leq k \leq 4 \times 10^4$ with $n = 10$ runs (*log scale*)

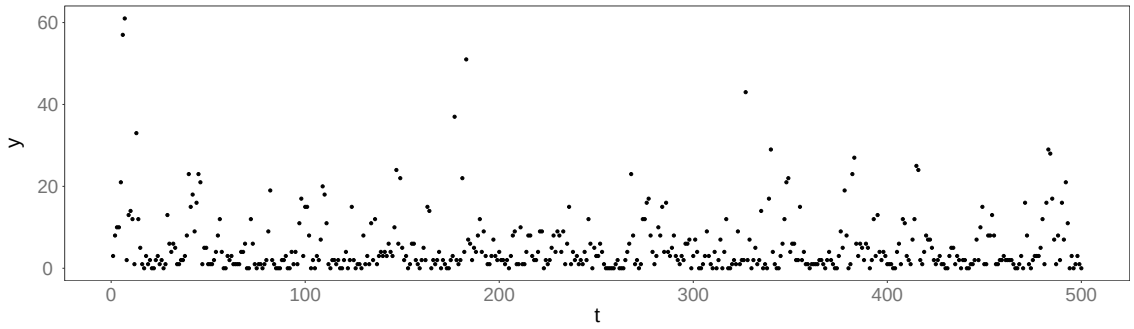


Figure 16.4: Realisation of an AR(1) Poisson DLM, $N_{obs} = 500$,$\boldsymbol{\Phi} = \{\alpha, \beta, \tau^2\} = \{0.5, 0.5, 1\}$.

For each SIR run the squared error, ϵ^2, was calculated for each time-point according to

$$\epsilon_t^2 = \left(\theta_t - \overline{\theta}_t\right)^2,$$

where θ_t represents the state posterior mean at time t estimated using a PMMH long run[2]. The ϵ^2 values for each resampling method can be viewed in Figure 16.5. The $\widehat{ESS}$ (as defined in (7.1)) for each run and resampler are also presented in Figure 16.5. In the summary in Table 16.1 we present the can $MSE_{\boldsymbol{\theta}}$ regarding the state estimation. Throughout this chapter we will use the $MSE_{\boldsymbol{\theta}}$ calculated as

$$MSE_{\boldsymbol{\theta}^i} = \frac{1}{N_{obs}} \sum_{t=1}^{N_{obs}} \left(\epsilon_t^i\right)^2,$$

where $\left(\epsilon_t^i\right)^2 = \left(\theta_t^i - \overline{\theta^i}_t\right)^2$ represent the squared error for the state vector i^{th} component.

From Table 16.1 we can see that stratified resampling performs slightly better, when using the $MSE_{\boldsymbol{\theta}}$ has a criteria, although all resampling methods produce consistent results. When taking into account the average $\overline{ESS}$ (averaged over $n = 500$ runs) the difference between the three resampling methods is negligible. Regarding computational times, we can see from Figure 16.1 that both stratified and systematic resampling have an advantage over multinomial resampling. These results are in accordance to the general findings in the literature (Douc *et al.* (2005)). Based on these results and the general advice in the literature, we have decided to use stratified resampling throughout this chapter (except when noted otherwise).

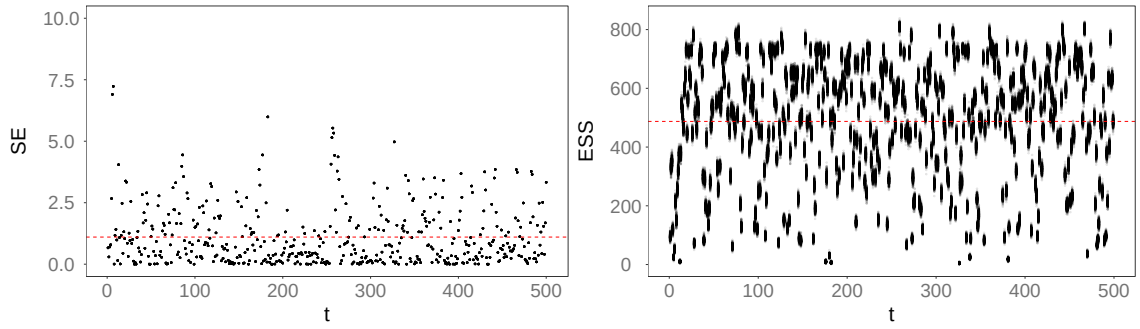
Regarding the frequency of the resampling step, as in previous examples, we have decided to use a static checkpoint of $n = 1$. Although it is well known that the resampling step theoretically increases the variance of the PF estimations (Chopin (2004); Creal (2012);

[2]Trace and ACF plots can be viewed in Appendix A.4

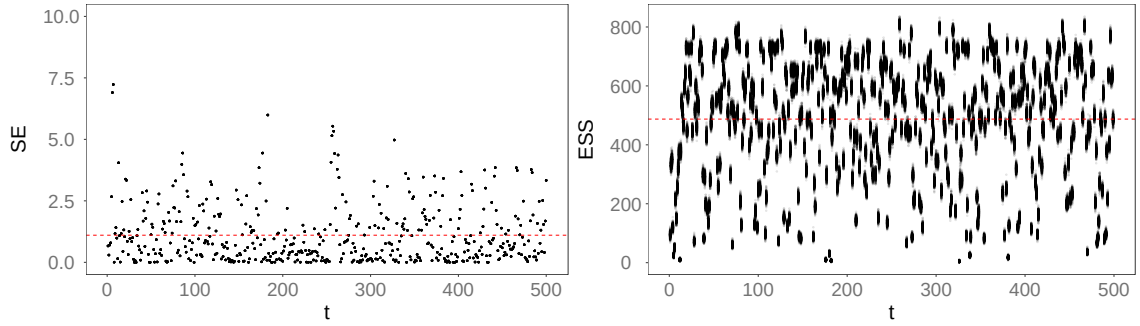
Resampler	$MSE_{\boldsymbol{\theta}}$	$\sigma_{MSE_{\boldsymbol{\theta}}}$	$\overline{ESS}$	$\sigma_{\overline{ESS}}$	**mean time (s)**	$\sigma_{\bar{t}}$
Multinomial	0.0185	10^{-4}	487.3524	0.0233	0.1387	0.001
Stratified	0.0182	10^{-4}	487.3739	0.0221	0.1196	0.001
Systematic	0.0183	10^{-4}	487.3255	0.0226	0.1134	0.001

Table 16.1: State estimation posterior mean MSE, average $\widehat{ESS}$ and computational times averaged for $n = 500$ runs (along with standard errors, σ) of a SIR filter on simulated data from a AR(1) PoDLM $\boldsymbol{\Phi} = \left\{\alpha, \beta, \tau^2\right\} = \{0.5, 0.5, 1\}$.

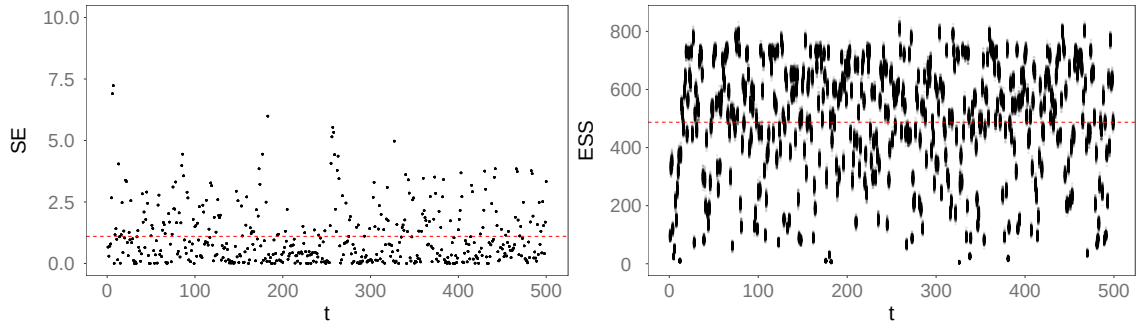
Hans (2003)) and introduce additional computational costs (although from Section 16.2 and specifically Figure 7.5 we have seen that this additional cost was in the order of ≈ 2 milliseconds at most for the resampling itself with $n = 4 \times 10^4$ weights), by performing resampling at each step we keep the computational cost strictly bounded (given that all other PF parameters are kept constant, such as N_p) in an online setting and evaluate the "worst case" scenario for near real-time streaming data settings.



(a) Multinomial



(b) Stratified



(c) Systematic

Figure 16.5: ϵ_t^2 (*left*) and $\widehat{ESS}$ (*right*) for $n = 500$ runs of a SIR filter on simulated data from a AR(1) PoDLM $\boldsymbol{\Phi} = \{\alpha, \beta, \tau^2\} = \{0.5, 0.5, 1\}$. Red line represents MSE (*left*) and mean $\widehat{ESS}$ (*right*).

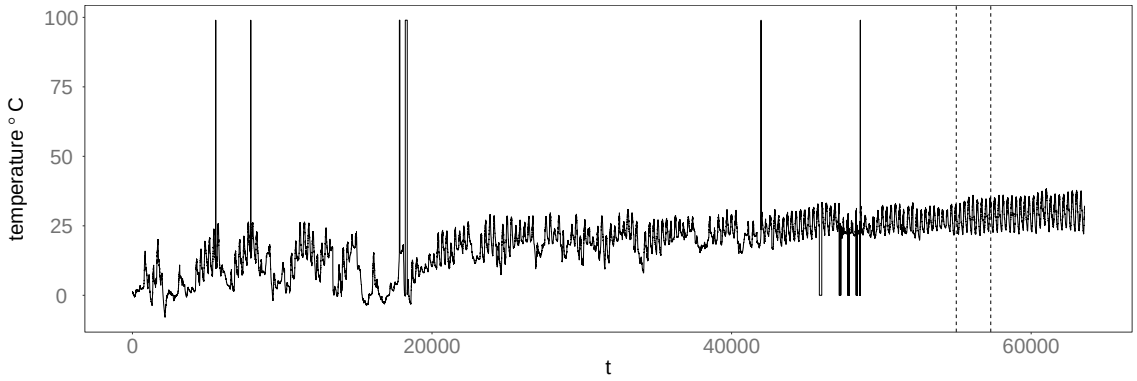


Figure 16.6: USCRN/USRCRN 2015 temperature data for Austin, Texas (USA)

16.3 Temperature data

The dataset used to test the Normal DLM implementation is provided by the U.S. Climate Reference Network / U.S. Regional Climate Reference Network (USCRN/USRCRN)[3]. The dataset used measures of air temperature in five minute intervals in the month of January 2015 at the city of Austin, Texas in the United States of America. This dataset exhibits a strong daily cyclical pattern.

Two separate analyses were performed with this dataset. One for a regular subset of the data, without visually identifiable outliers, from $t = 55000$ to $t = 57303$, that is approximately 8 days worth of observations in Section 16.3.1. The second analysis, in Section 16.3.2 for a subset of the data with the presence of outliers, with 4 days worth of data, from $47870 \leq t \leq 49022$. Although it would be easy to identify and strip the erroneous outlier value in this case, we deliberately leave them in order to test the robustness of our filtering methods to rogue values inconsistent with the model.

16.3.1 Temperature dataset A

We decided to model this data as a Normal DLM following the relations of (2.17) and (2.18). The complete dataset clearly includes seasonal components contributing with different periods, such as daily and yearly seasonality. However, since we applied the estimation to a subset corresponding to a week's worth of measurements, we decided to ignore the yearly and monthly effects as within our time window these effects could be captured by an underlying mean.

A locally constant component ($\mathcal{P}(1)$) was incorporated, to capture the underlying mean, and a daily Fourier component ($\mathcal{F}(\cdot)$) was also included to account for the daily

[3]Available at: http://www1.ncdc.noaa.gov/pub/data/uscrn/products/subhourly01/ (Accessed 13^{th} September 2017)

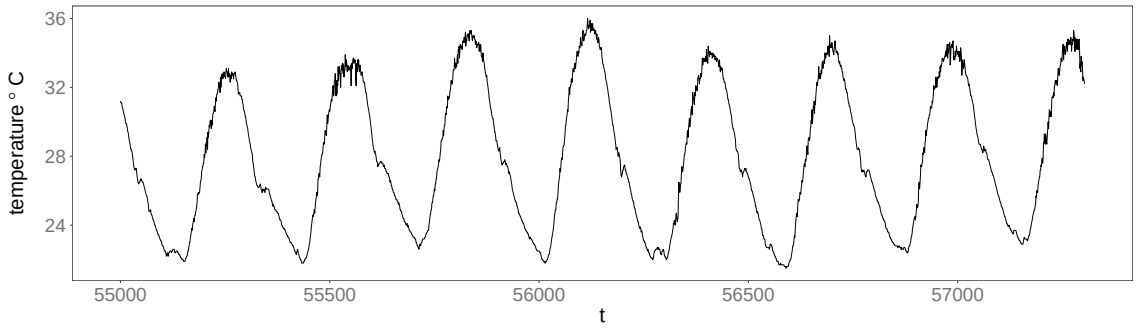


Figure 16.7: USCRN/USRCRN 2015 temperature data for Austin, Texas (USA) for $55000 \leq t \leq 57303$.

cyclical effect. The model will then consist of a locally constant component, with variance τ_μ^2 and prior $\tau_\mu^2 \sim \mathcal{IG}(1,1)$, and a daily seasonal Fourier component. As the data is sampled every 5 minutes, the period of the Fourier components will be $p = 288$. We will use three harmonics ($\mathcal{F}(288,3)$) and therefore, according to (2.15), the diagonal of our covariance matrix will be $\text{diag}\left(\mathsf{W}^{daily}\right) = \left(\tau_{d,1}^2, \tau_{d,2}^2, \tau_{d,3}^2, \tau_{d,4}^2, \tau_{d,5}^2, \tau_{d,6}^2, \tau_{d,7}^2\right)$. The prior was set to $p\left(\mathsf{W}^{daily}\right) = \mathcal{IW}(7, \mathbf{I})$. This is equivalent to the formulation

$$
\begin{aligned}
y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\mathsf{F}^T\boldsymbol{\theta}_t, V\right) \\
\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}\right)
\end{aligned}
$$

with structural matrices

$$
\mathsf{F} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}^T,
$$

$$
\mathsf{G} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \cos\frac{2\pi}{p} & \sin\frac{2\pi}{p} & 0 & 0 & 0 & 0 \\
0 & -\sin\frac{2\pi}{p} & \cos\frac{2\pi}{p} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \cos\frac{4\pi}{p} & \sin\frac{4\pi}{p} & 0 & 0 \\
0 & 0 & 0 & -\sin\frac{4\pi}{p} & \cos\frac{4\pi}{p} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \cos\frac{6\pi}{p} & \sin\frac{6\pi}{p} \\
0 & 0 & 0 & 0 & 0 & -\sin\frac{6\pi}{p} & \cos\frac{6\pi}{p}
\end{bmatrix},
$$

$p = 288$.

As show in section 2.2.2.1 on page 9 each additional h harmonics will lead to an increase of $2h$ in the number of parameters and size of the state vector. This will directly grow

the computational cost of the SMC methods by increasing computational complexity, but also indirectly, since the number of particles would also have to increase in order to better explore the state and parameter space. A run of the Storvik filter (with $N_p = 3 \times 10^4$) was performed for $h = 1, 2, \ldots, 5$ and the observation one-step ahead forecast was used to calculate the MSE between y_{t+1} and $\widehat{y}_{t+1}$ for $t = 1, \ldots, N_{obs} - 1$) in order to assess forecast performance. We can see from Figure C.1 and Table C.1 that there is a considerable increase in forecast performance up until $h = 3$, after which the MSE doesn't reduce as much. However, the total time of each run grows linearly with the number of harmonics. As such, the number of harmonics chosen, $h = 3$, is a compromise between having an approximation of the seasonal component which is more flexible than the simplest case ($h = 1$) and a need to keep the model size, and consequently, the computational cost, suitable for near real-time inference.

16.3.1.1 Offline estimation

Offline estimation of the parameter set $\boldsymbol{\Phi} = \{\mathsf{W}, V\}$ was performed using PMMH, IBIS, O-IBIS and EM[4]. The IBIS and O-IBIS were also implemented in the variants IBIS-SVD and O-IBIS-SVD, where the associated KF recursions used the SVD formulation as described in Section 4.1 on page 36. Since a Normal DLM is used for the temperature data, it is possible to use a MH scheme which makes use of the exact likelihood provided by the KF recursions. Although we have used that ability in this section (with IBIS/SVD, O-IBIS/SVD and the fully adapted proposals for the online filters), PMMH was used as the "gold standard" for all the datasets for consistency. The marginal parameter posterior using PMMH, IBIS/SVD and O-IBIS/SVD at time $t = N_{obs}$ is presented in Figure 16.8 on page 200. The estimation history for IBIS/SVD, O-IBIS/SVD and EM are presented respectively in Figures 16.9 on page 201 and 16.10 on page 202. The traces and auto-correlation plots for PMMH can be viewed in the Appendix A.5.1 on page 281.

IBIS/SVD and O-IBIS/SVD were both performed with a number of particles $N_p = 2 \times 10^4$ and in the O-IBIS/SVD case the observation window was $h = 500$ observations, both using state priors $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{0}, 100\mathbf{I}_7)$. The resampler used for IBIS and O-IBIS was the systematic resampler and the resampling-move step occurred at $\widehat{ESS} < N_p/2$. PMMH used a bootstrap filter with $N_p = 2000$ with a systematic resampler at static checkpoint $n = 1$. The priors used with IBIS/SVD, O-IBIS/SVD and PMMH for $\boldsymbol{\Phi}_0$ were

$$\tau_\mu^2, \tau_{d,1:7}^2 \sim \mathcal{IG}(1, 1), \qquad V = \sigma^2 \sim \mathcal{IG}(1, 1).$$

A summary of parameter estimation results for the offline methods is presented in Table 16.2.

[4]Used to calculate the MLE of $\boldsymbol{\Phi}$.

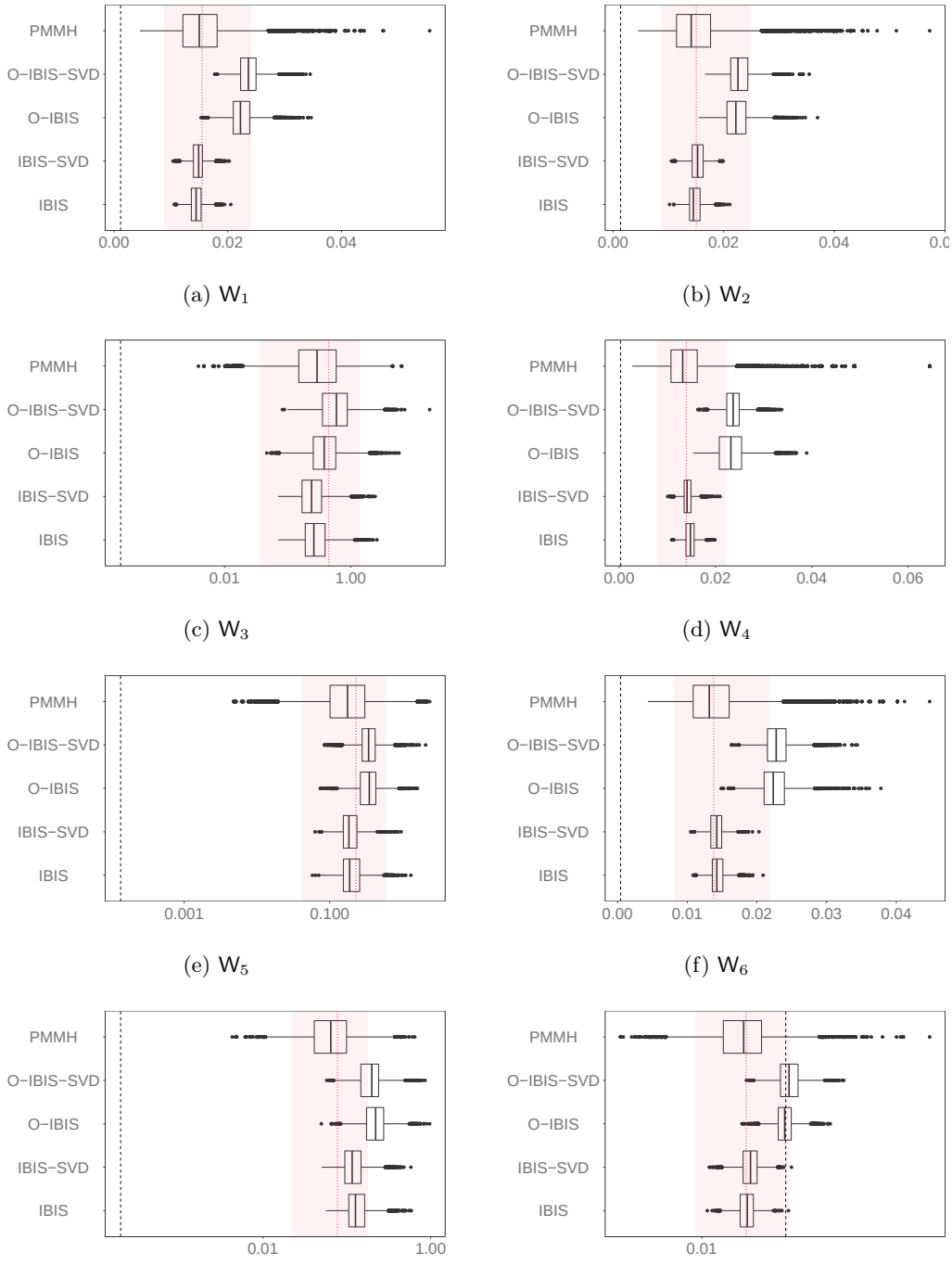
Method	**Time (s)**	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$	$\overline{\mathsf{W}}_4$	$\overline{\mathsf{W}}_5$	$\overline{\mathsf{W}}_6$	$\overline{\mathsf{W}}_7$	$\overline{V}$
IBIS	2440.49	0.0145	0.0147	0.3226	0.0147	0.2160	0.0143	0.1407	0.0145
IBIS-SVD	3786.65	0.0148	0.0152	0.2920	0.0142	0.2036	0.0142	0.1280	0.0148
O-IBIS	1243.45	0.0225	0.0224	0.7409	0.0233	0.3727	0.0225	0.2336	0.0198
O-IBIS-SVD	3099.44	0.0237	0.0228	0.7230	0.0236	0.3657	0.0229	0.2084	0.0206
EM[5]	42.65	0.0011	0.0013	0.0002	0.0002	0.0001	0.0004	0.0002	0.0199
PMMH	–	0.0155	0.0150	0.4481	0.0139	0.2295	0.0138	0.0774	0.0144

Table 16.2: Parameter posterior mean estimation and computation time with offline (including O-IBIS) methods for the temperature dataset A.

The computational times for each *resampling* step of IBIS and O-IBIS are presented in Figure 16.11.

The initial $\boldsymbol{\Phi}_0$ for the EM estimation was the same for all parameters and chosen as an arbitrary value of $\boldsymbol{\Phi}_0 = 10$ and the tolerance to test convergence was set at $\epsilon = 10^{-11}$. The EM reached the convergence criterion in 30975 iterations taking 42.65 seconds.

We can see from Figure 16.9 that both the IBIS and O-IBIS parameter estimation come close to the PMMH estimated values. Although the estimation provided by "full" IBIS (that is, using the entirety of the data up until time t) is closer to the PMMH value, there is overlap present between IBIS and O-IBIS. We can also observe that O-IBIS tends to over estimate the variance of the posterior (as evident in, for instance the posterior for W_6). This behaviour was already noted in Section 14.1 on page 171 where as the observation window h increases, the O-IBIS posterior estimation at time t will approximate the "full" IBIS. When comparing the parameter estimation of EM with PMMH, IBIS/SVD and O-IBIS/SVD, we can see (Table 16.2) that the IBIS based methods are closer to the PMMH results than the EM results by several orders of magnitude. Looking at the EM convergence plots in Figure 16.10, we can verify that this was not simply a case of insufficient iterations, since the parameter estimation seems to stabilise quite early on in the iterative process.



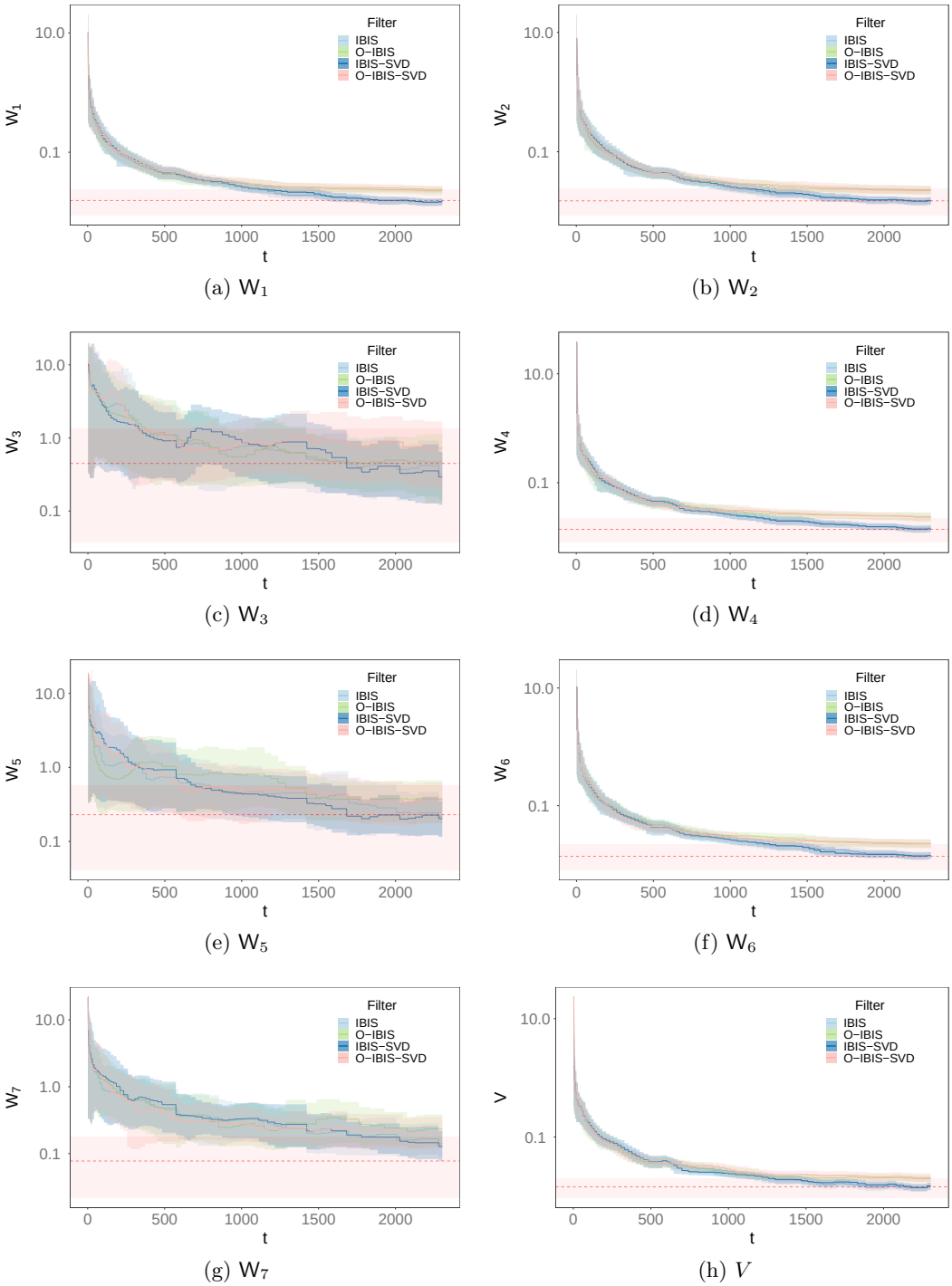
(a) W_1 (b) W_2

(c) W_3 (d) W_4

(e) W_5 (f) W_6

(g) W_7 (h) V

Figure 16.8: Marginal (*log-scale*) for parameters $\boldsymbol{\Phi} = \{\mathsf{W}, V\}$ at $t = N_{obs}$ using PMMH, IBIS/SVD and O-IBIS/SVD on the temperature dataset A. Vertical black line is the EM estimation. Vertical red line is the PMMH posterior mean and shaded area 90% equitailed credibility interval.



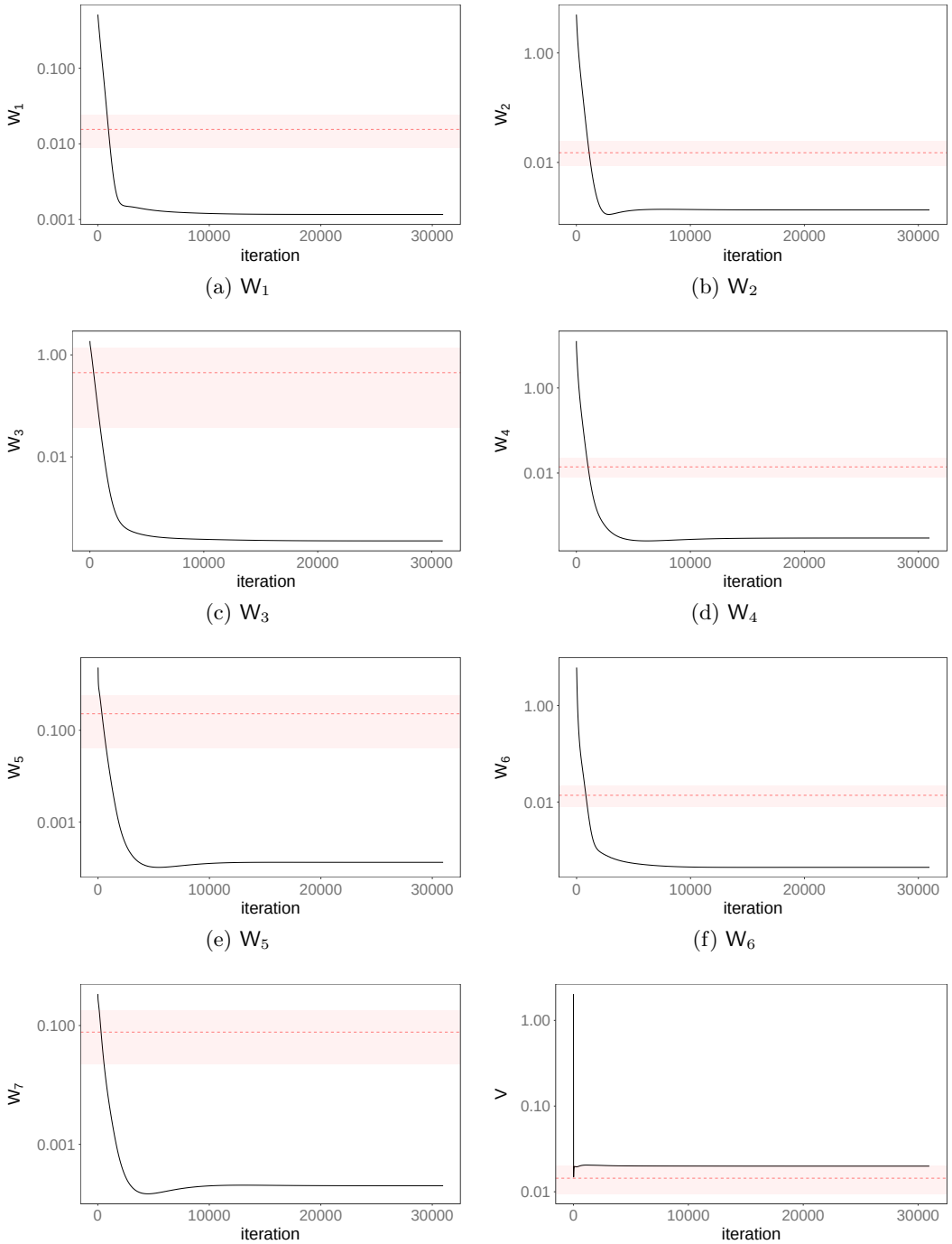
(a) W_1 (b) W_2

(c) W_3 (d) W_4

(e) W_5 (f) W_6

(g) W_7 (h) V

Figure 16.9: Estimation history for parameters $\boldsymbol{\Phi} = \{W, V\}$ for the temperature dataset A. Solid lines represent the posterior mean and shaded areas the 90% equitailed credibility interval. PMMH in red (dashed horizontal line is the posterior mean and shaded area the 90% quantile interval). using IBIS/SVD and O-IBIS/SVD.



(a) W_1 (b) W_2

(c) W_3 (d) W_4

(e) W_5 (f) W_6

(g) W_7 (h) V

Figure 16.10: Estimation history (log scale) for parameters $\boldsymbol{\Phi} = \{\mathsf{W}, V\}$ using EM for the temperature dataset. Horizontal dashed line represents the PMMH posterior's average and pink band the 90% equitailed credibility interval.

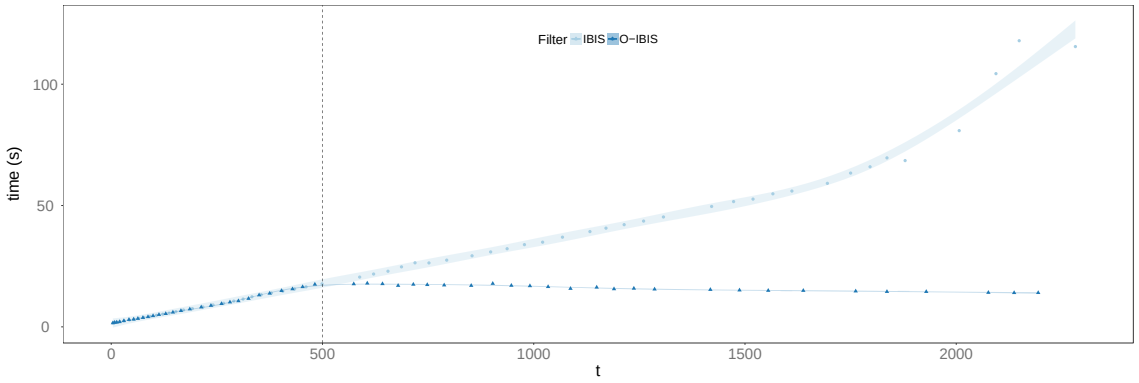


Figure 16.11: Computational time at each resampling step for IBIS and O-IBIS with the temperature dataset A

Method	**Time (s)**	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$	$\overline{\mathsf{W}}_4$	$\overline{\mathsf{W}}_5$	$\overline{\mathsf{W}}_6$	$\overline{\mathsf{W}}_7$	$\overline{V}$
LW	866.923	0.1073	0.1320	2.4023	0.0901	2.3045	0.0859	1.3657	0.0881
Storvik	1566.855	0.0653	0.0642	1.3028	0.0841	0.8667	0.0693	0.5467	0.0427
PL	1684.522	0.0530	0.0631	0.3063	0.0530	0.3775	0.0495	0.4919	0.0393
PMMH	–	0.0155	0.0150	0.4481	0.0139	0.2295	0.0138	0.0774	0.0144
		σ_{W_1}	σ_{W_2}	σ_{W_3}	σ_{W_4}	σ_{W_5}	σ_{W_6}	σ_{W_7}	σ_V
LW		0.0033	0.0039	0.1636	0.0035	0.1706	0.003	0.0658	0.0021
Storvik		0.0011	0.0006	0.0074	0.0009	0.0075	0.0006	0.004	0.0005
PL		0.0005	0.0004	0.0061	0.0006	0.004	0.0006	0.003	0.0006
PMMH		0.0048	0.0051	0.4582	0.0047	0.1917	0.0043	0.0525	0.0035

Table 16.3: Parameter posterior mean (and standard deviation) estimation at $t = N_{obs}$ with online methods for the temperature dataset A.

16.3.1.2 Online estimation

Online state and parameter estimation was performed using the Liu & West, Storvik and Particle Learning filters. The model structure used was the same as the one presented in Section 16.3.1 on page 196. The state priors were the same for all the filters and drawn from a vague prior $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0)$, with the mean component approximately centred around the average yearly temperature range of 15°C, such that

$$\mathbf{m}_0 = [15 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T,$$
$$\mathsf{C}_0 = \text{diag}\,(400, 100, 100, 100, 100, 100, 100)\,.$$

[5] The EM method does not provide with a posterior for $\boldsymbol{\Phi}$. The value corresponds to the point estimate at final iteration.

The number of particles was chosen after testing[6] the Storvik filter (LW was excluded from the test on account of the extra parameter to control for, namely the smoothing parameter δ) with a selected range of values from $100 \leq N_p \leq 3 \times 10^4$. After performing $n = 50$ runs, both the parameter (Figure D.1) and state (Figure D.2) posterior means at $t = N_{obs}$ show small variability between different values of N_p for $N_p \geq 2 \times 10^4$ and are generally consistent with the PMMH posterior mean estimation. For this reason, and also taking into consideration the computational cost, the value of $N_p = 3 \times 10^4$ was chosen. The shrinkage factor used for Liu & West was $\delta = 0.99$ and the resampler used was the stratified resampler presented in (7.3.3) with a static checkpoint of $n = 1$, that is, performing resampling at every time step t. In order to choose the δ for the Liu & West filter several runs were performed[7] for different values, namely $\delta = 0.90, 0.91, \ldots, 0.99$. The parameter priors used were

$$\begin{aligned} \sigma_0^2 &\sim \mathcal{IG}(1, 1) \\ \mathsf{W}_0 &\sim \mathcal{IW}(7, 0.2\mathbf{I}_7). \end{aligned}$$

The fully adapted version of the filters was used, according to the specifications of Section 6.2.1 on page 73. The importance density was therefore the optimal importance density. We will present the estimation for approximately 1 week's worth of data ($N_{obs} = 2304$).

The filter's results are compared to a long run of a PMMH and MSE is calculated considering the PMMH's result as the "true" value. In Figure 16.12 on page 206 we present the estimated marginals for the parameter set $\boldsymbol{\Phi} = \{\mathsf{W}, V\}$ using the online methods, in comparison to the offline results obtained in Section 16.3.1.1. In Figure 16.13 on page 207 we show the estimation history for $\boldsymbol{\Phi}$ using the online methods and a summary of the online parameter estimation results is presented in Table 16.3 on the preceding page.

The results for online state estimation is presented separately for each state component $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_7\}$ in Figure 16.14 on page 208, along with the ESS for each of the online methods. A summary of the MSE between each state component and the PMMH state estimation is presented in Table 16.4 on the facing page, along with each filter's total and iteration execution times.

We can verify from Table 16.2 and Figure 16.12 that the sufficient statistic based methods (Storvik and PL) approximate the parameter set more accurately (relatively to the PMMH estimate), but still not as accurately as the IBIS/SVD implementation. An important factor in this analysis within the context of near real-time state and parameter

[6] A summary of the tuning results can be found in Appendix D.1.
[7] A summary of the tuning results can be found in Appendix E.1.1.

Method	MSE							Time	
	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	$\boldsymbol{\theta}_4$	$\boldsymbol{\theta}_5$	$\boldsymbol{\theta}_6$	$\boldsymbol{\theta}_7$	*iteration (ms)*	*total (s)*
LW	5.6362	4.8364	1.8079	1.2451	1.383	0.8827	0.6062	376.3	866.92
Storvik	7.8953	4.7619	2.7269	1.8106	1.416	0.7538	0.5177	680.1	1566.86
PL	2.2519	2.4766	0.8508	0.3	0.564	0.4244	0.3609	731.1	1684.52

Table 16.4: MSE for different particle filters with the temperature dataset A compared to the PMMH estimation and computational time for $N_p = 3 \times 10^4, N_{obs} = 2304$.

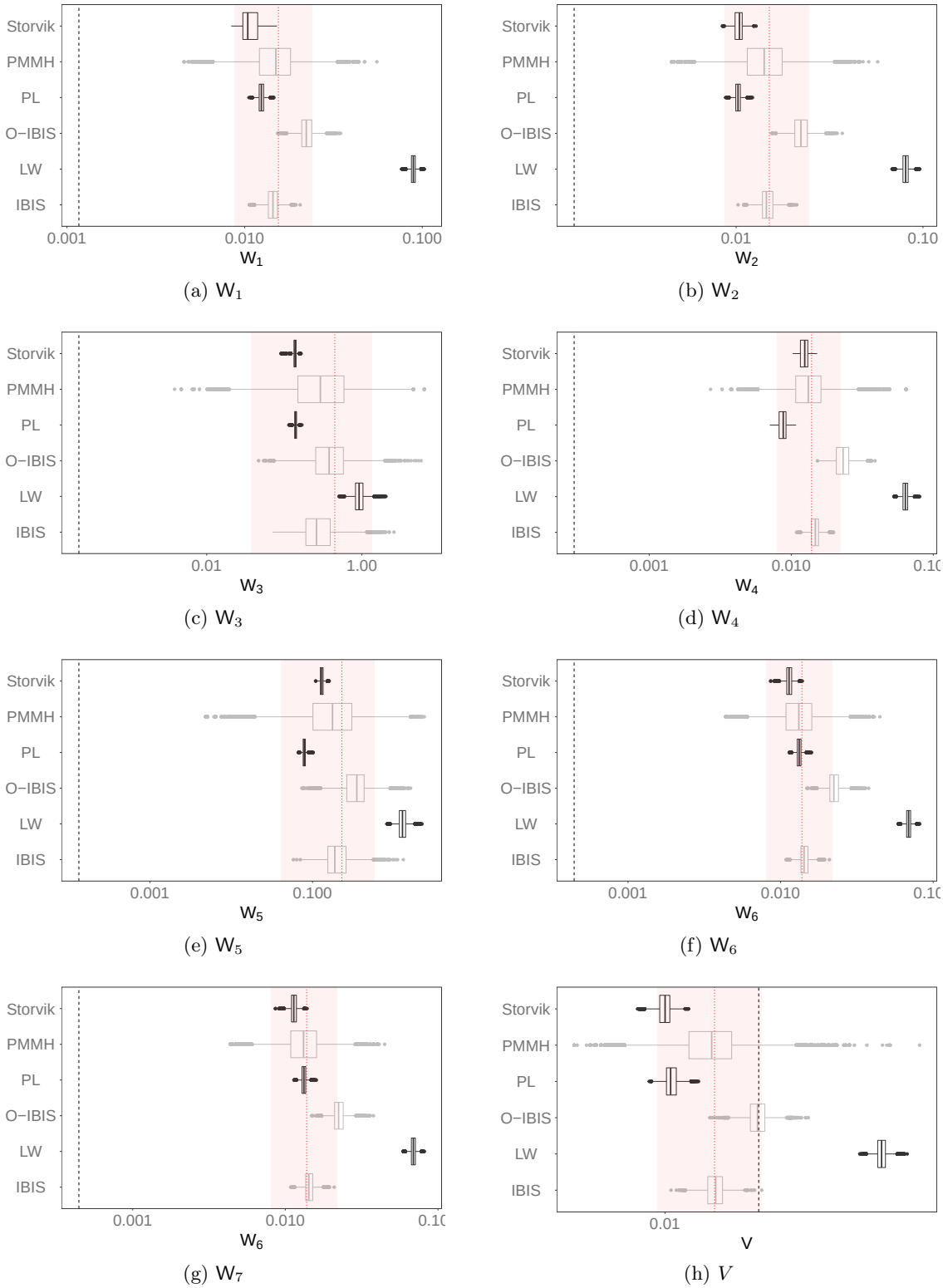
estimation in streaming data is obviously the computational cost. We can see in Table 16.4 that even with a substantial amount of particles, $N_p = 3 \times 10^4$, the execution time of each step is within the range of 376.3 to 731.1 milliseconds (for LW and PL, respectively). This computational cost is bound and $\mathcal{O}(N_p)$ conditioned on the model's dimension, that is, each step will only increase its cost per iteration for the same model linearly as a function of the number of particles used. This cost is perfectly reasonable for near real-time estimation on streaming data at incoming rates of one observation per second, for instance.

In the case of O-IBIS-SVD, there are different considerations to be taken into account regarding computational cost. One of them is obviously the resampling step frequency (since the particle rejuvenation step is the actual computational bottleneck). However, even when considering the extreme case of static checkpoints with $n = 1$, *i.e.* resample-move at every time t, the computational cost will also be bounded, but in this case with $\mathcal{O}(hN_{\boldsymbol{\Phi}})$, that is, conditioned on the model's dimension, the computational cost will be constant for a given window size, h and the number of parameter-particles, and also computationally bound nevertheless making it an "online" method.

In terms of state estimation, we can see from Table 16.4 that PL outperforms both Storvik and LW when using the state's MSE relatively to the PMMH as a criteria. Also noteworthy is the fact that state estimation's marginal posterior, $\hat{p}(\boldsymbol{\theta}_t|\mathcal{D}_t)$ variance is smaller for the PL than the other filters and generally smaller for sufficient statistics based methods when compared to LW. The ESS is however consistent between the three online methods (with sufficient statistics-based methods having marginally higher values) and this is a behaviour we observe mainly when using the optimal importance density.

16.3.1.3 Forecast

The results for the one-step ahead observation forecast and respective errors, $e_t = (y_t - \hat{y}_t)$ are presented in Figures 16.15 on page 210 and 16.18. A summary of the observation one-step ahead MSE is presented in Table 16.5 on page 209.



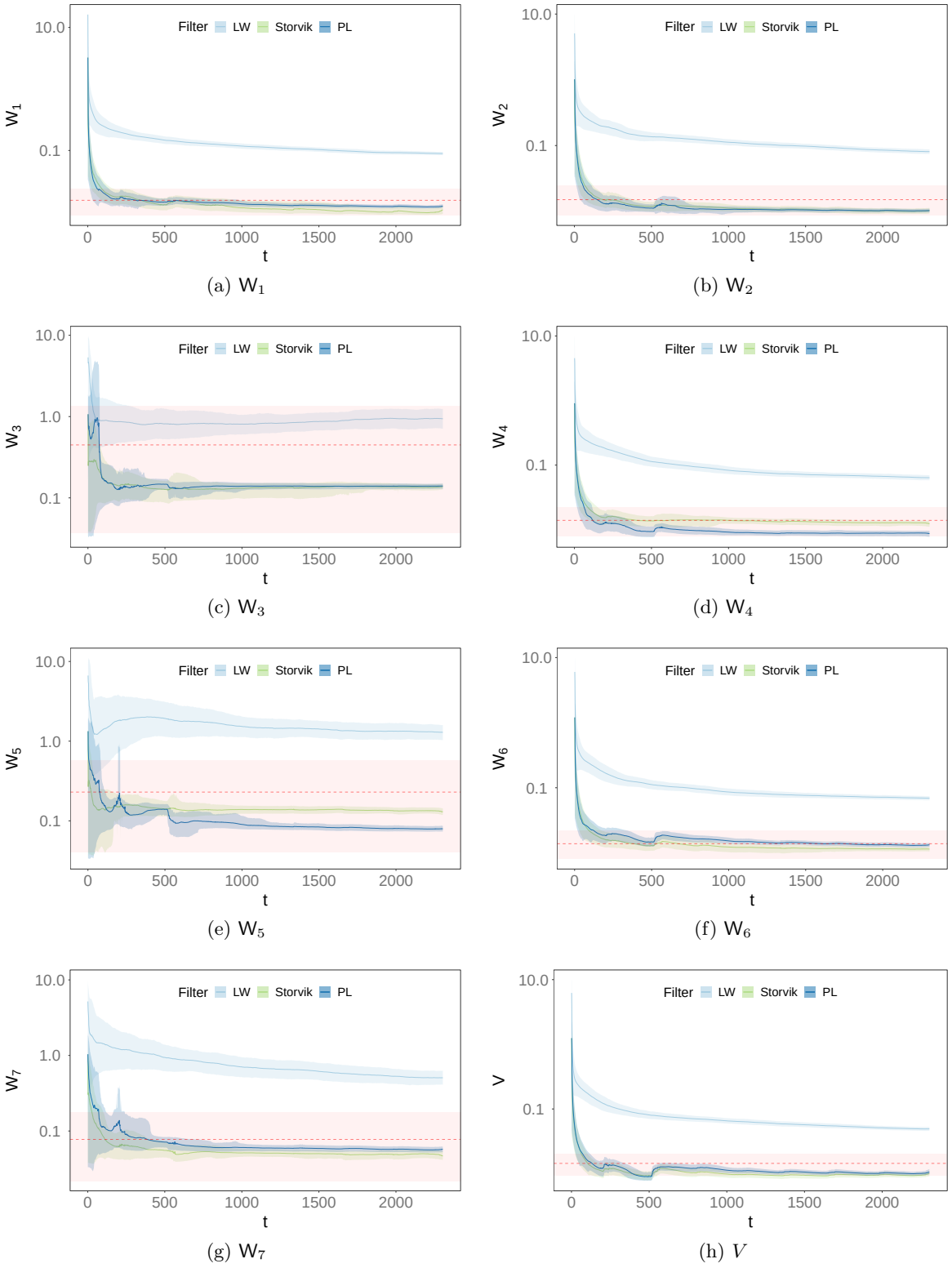
(a) W_1 (b) W_2

(c) W_3 (d) W_4

(e) W_5 (f) W_6

(g) W_7 (h) V

Figure 16.12: Estimated posterior for $\boldsymbol{\Phi}$ at $t = N_{obs}$ for the online methods, compared to the offline methods for temperature dataset A. Vertical black line represents the EM estimation, red dashed line the PMMH mean and vertical red band the 90% equitailed credibility interval.



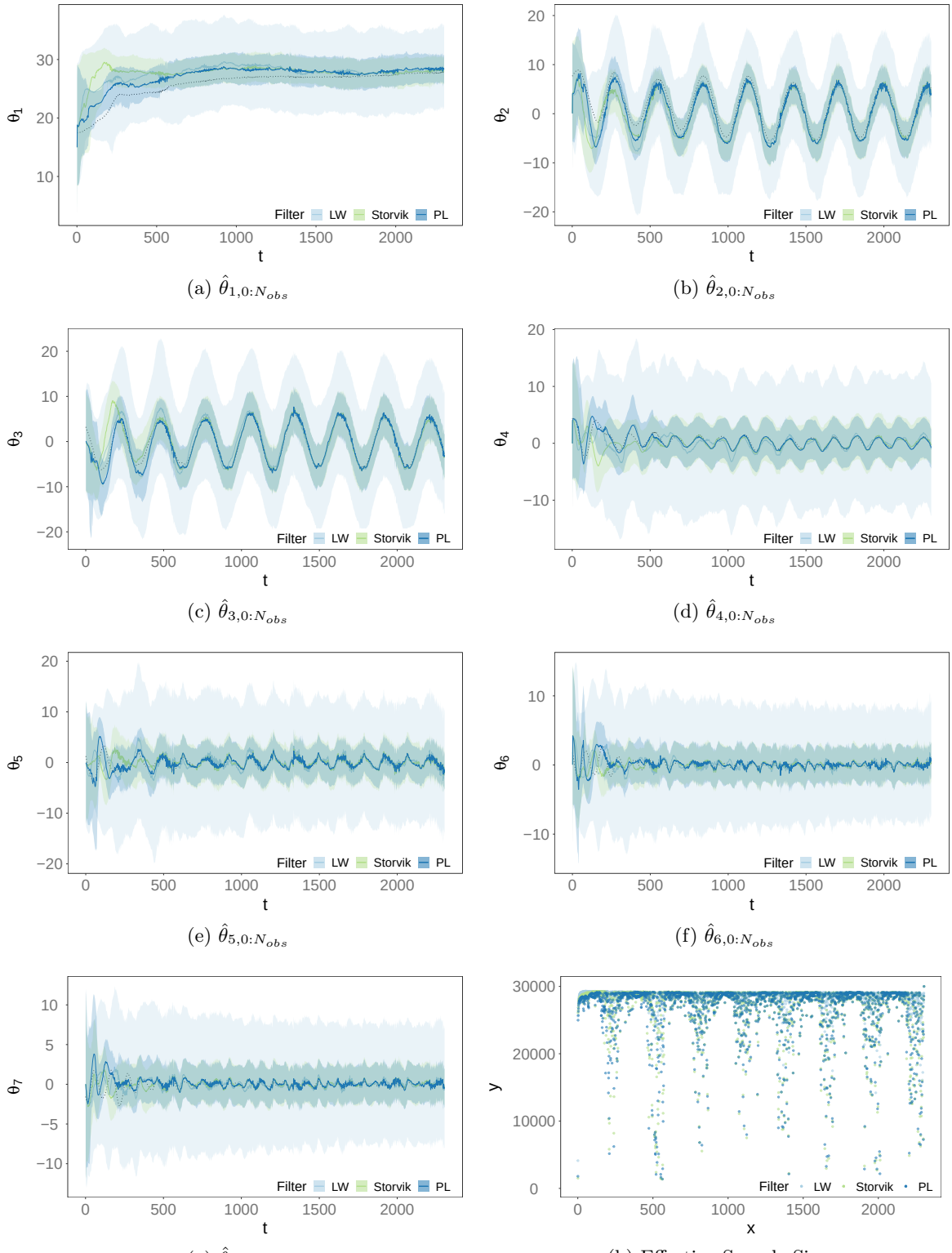
(a) W_1 (b) W_2

(c) W_3 (d) W_4

(e) W_5 (f) W_6

(g) W_7 (h) V

Figure 16.13: Estimation history for $\boldsymbol{\Phi}$ using the online methods for the temperature dataset A. Horizontal line represents the PMMH posterior mean and colour lines the PFs estimated posterior mean. Shaded areas represent 90% equitailed credibility interval.



(a) $\hat{\theta}_{1,0:N_{obs}}$

(b) $\hat{\theta}_{2,0:N_{obs}}$

(c) $\hat{\theta}_{3,0:N_{obs}}$

(d) $\hat{\theta}_{4,0:N_{obs}}$

(e) $\hat{\theta}_{5,0:N_{obs}}$

(f) $\hat{\theta}_{6,0:N_{obs}}$

(g) $\hat{\theta}_{7,0:N_{obs}}$

(h) Effective Sample Size

Figure 16.14: State components for PF estimation on the temperature dataset A ($N_{obs} = 2034$) and $\widehat{ESS}$ (*bottom left*). Colour lines represent the posterior mean and shaded areas the 90% equitailed credibility interval. Dashed black line represents the PMMH state posterior mean.

Method	MSE								
	$\widehat{y}_{t+1}$	$\widehat{y}_{t+k}$	$\widehat{\boldsymbol{\theta}}_{1,t+k}$	$\widehat{\boldsymbol{\theta}}_{2,t+k}$	$\widehat{\boldsymbol{\theta}}_{3,t+k}$	$\widehat{\boldsymbol{\theta}}_{4,t+k}$	$\widehat{\boldsymbol{\theta}}_{5,t+k}$	$\widehat{\boldsymbol{\theta}}_{6,t+k}$	$\widehat{\boldsymbol{\theta}}_{7,t+k}$
LW	0.00056	0.9748	0.3623	0.7898	0.3099	0.0356	0.0354	0.3119	0.3226
Storvik	0.00055	0.8193	0.5176	0.6601	0.2179	0.085	0.0864	0.1896	0.2
PL	0.00055	0.7659	0.6369	0.6229	0.1813	0.1336	0.1335	0.1537	0.1661

Table 16.5: One-step and k-step ($k = 804$) ahead forecast Mean Squared Error (MSE) for different particle filters with the temperature dataset A.

A longer k-step ahead forecast, for $k = 804$, approximately 2 days and 19 hours worth of temperature data was performed. The k-step ahead state forecast is represented in Figure 16.17 on page 212, for each individual component. A $k = 804$ step ahead forecast for the observations was equally performed with the result presented in Figure 16.16, along with the squared errors between the temperature forecast and the observed values. The summary of the k-step ahead forecast is also presented in Table 16.5.

We can see from Table 16.5 that the one-step ahead forecast MSE relatively to the true y_t has virtually identical values for the three online filters considered with slightly more accurate values for Storvik and PL. When considering long term observation forecasts, PL has a clear advantage producing a more accurate forecast when compared to the true observations $\mathcal{D}_{(T-k:T)}$ but only with a marginally smaller forecast variance when compared to the other methods. The long term observation forecast can be seen in Figure 16.16 and the squared errors ϵ_t^2 can be viewed in the same figure, reinforcing the sufficient-statistics based methods superior accuracy.

When looking at long-term state forecasts, PL has an advantage when compared to the remaining methods. We can also verify this visually when looking at the state forecast plots in Figure 16.17. PL tends to be much more consistent with the PMMH estimation than the remaining methods while showing a significantly smaller forecast variance than LW.

16.3.1.4 Monte Carlo variance

For the calculation of the $MCMAE$ for the parameters, LW, Storvik, PL and O-IBIS-SVD were used with state and parameter priors as specified in Section 16.3.1.1 whilst the number of particles was $N_p = 5000$. O-IBIS-SVD was used with $N_p = 1000$ and and observation window $h = 500$. The results were then averaged over $n = 50$ runs in order to obtain $\overline{MCMAE}$. These results are illustrated in Figure 16.19 and summarised in Table 16.16. Regarding the variability of the state estimation, the mean $MCMAE$ was also calculated

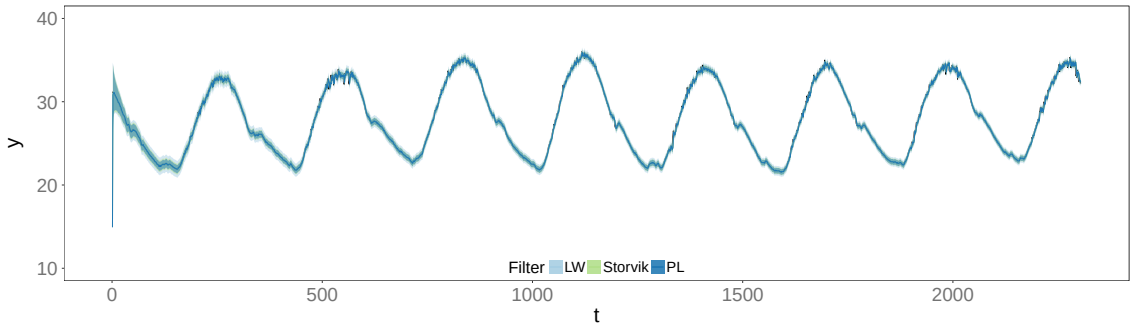


Figure 16.15: One-step ahead forecast for the online filters with the temperature data

(and presented in Figure 16.20). As in the parameter case, the mean $MCMAE$ will then be the averaged value for n runs. We also look at the parameter posterior mean and state posterior mean average (and standard errors) at $t = N_{obs}$ for the online filters and compare it to the PMMH state and parameter posterior estimation at $t = N_{obs}$. These results[8] are summarised, respectively, in Tables 16.8 on page 213 and 16.7 on page 213.

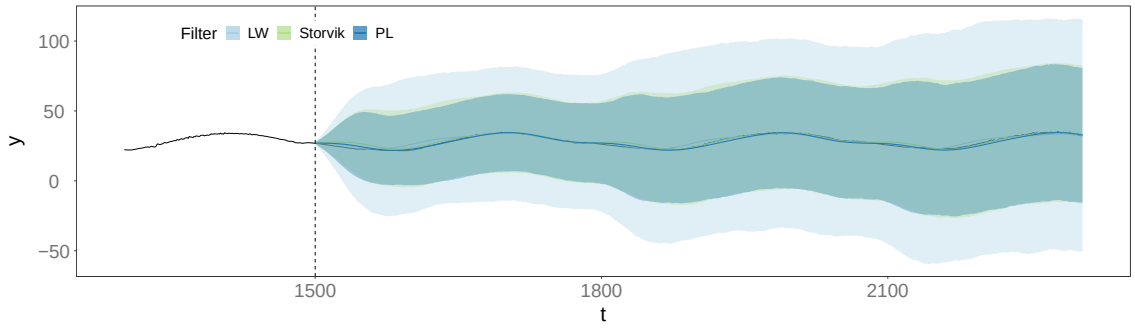
For Table 16.6 and Figure 16.19 we can clearly see that regarding parameter estimation, sufficient statistic-based methods (especially PL) display a much lower variability between runs than the LW filter. However, O-IBIS-SVD, is the more stable of all analysed methods, consistently converging to the PMMH estimated parameter for all runs. This is clearly visible in in Figure 16.19 for parameter W_7, the highest frequency harmonic of the component $\mathcal{F}(288, 3)$. PL, Storvik and LW seem to settle on very different values after a few iterations (perhaps suggesting an identifiability problem), whereas O-IBIS-SVD consistently converges to the PMMH estimation for all n runs.

In terms of state estimation variability, we can see that once more, sufficient-statistics-based methods display less variability from Table 16.6, with PL being the less variable. We can visually inspect this behaviour in Figure 16.20, where the we can also observe the effect of convergence to the PMMH values in latter stages of the state estimation when compared to the initial time steps.

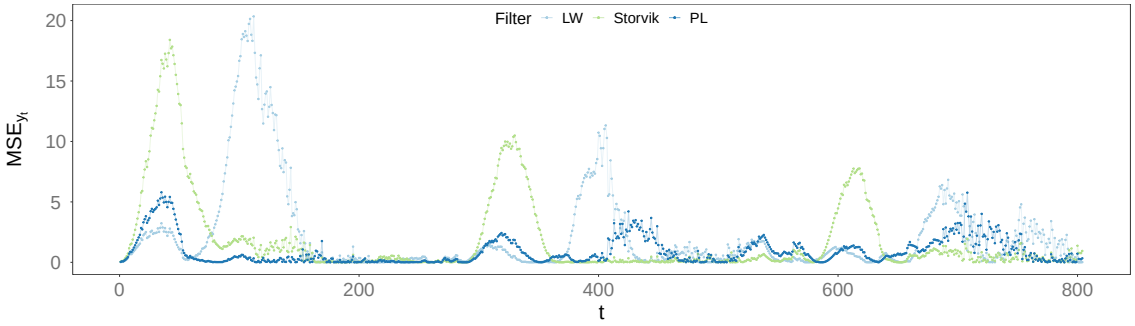
16.3.1.5 Discrepancy

One of the properties, as discussed in Section 2.4.3 on page 22 about the online estimation methods is the ability to calculate a discrepancy value d_t which would enable us to detect potential anomalies or outliers in the data stream. As we can see from Figure 16.21, no anomalies were detected with any of the online filters, when using a threshold of $d = 3$. This was to be expected, since as stated in the beginning of this section, the temperature

[8] Average posterior means can be viewed in Figures B.1 and B.2, in Appendix B.1.



(a) y_{t+k} forecast for the temperature data

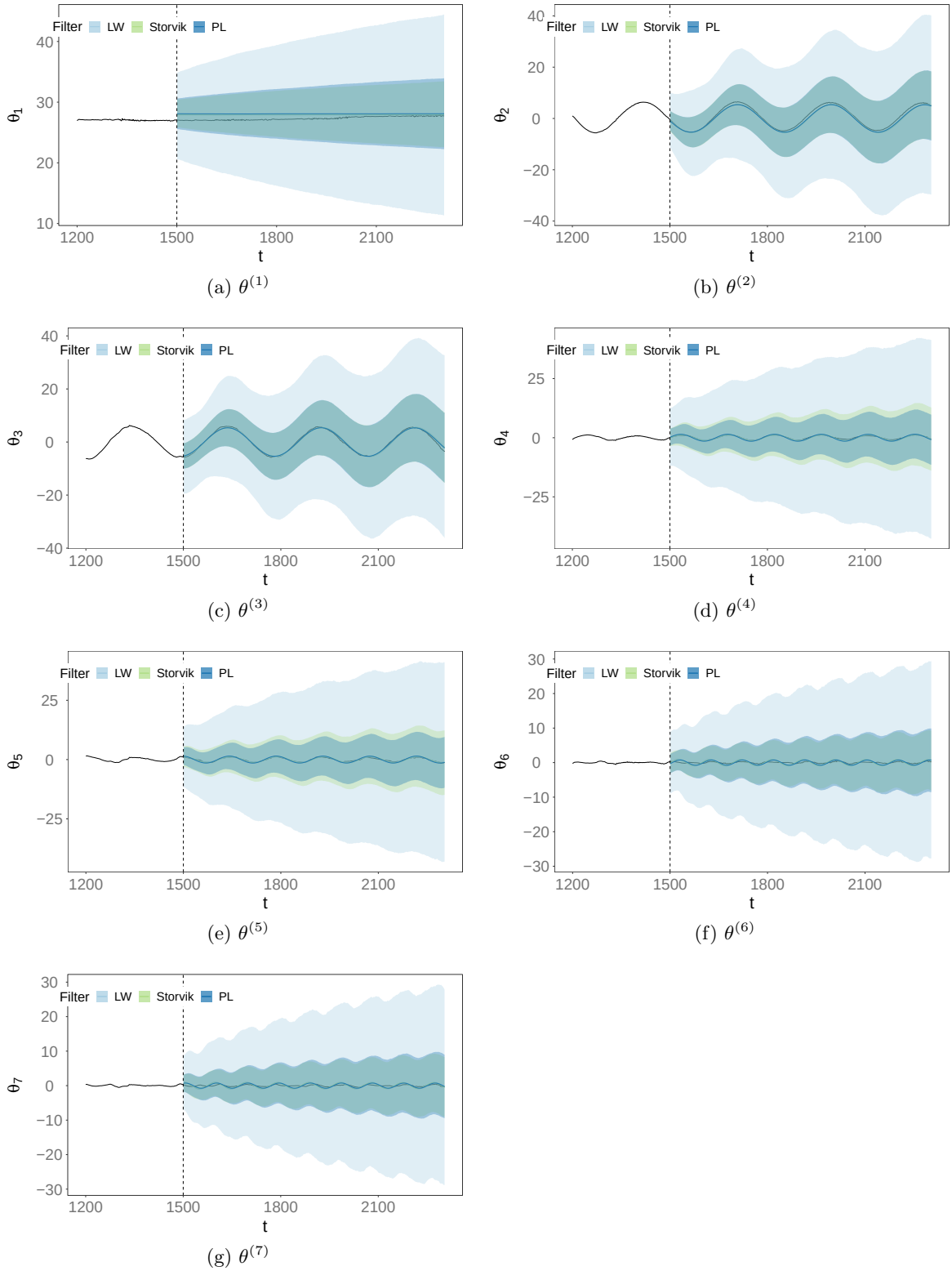


(b) MSE for y_{t+k} forecast

Figure 16.16: Observation forecast values (colour lines are the forecast density mean and shaded areas the 90% equitailed credibility interval) and MSE for different particles with the temperature dataset A.

Method	$\overline{MCMAE}$							
	W_1	W_2	W_3	W_4	W_5	W_6	W_7	V
LW	0.2755	0.2548	4.3692	0.2595	2.3225	0.2471	1.6637	0.1812
Storvik	0.1120	0.1117	1.2924	0.1126	1.4822	0.1102	0.9145	0.0730
PL	0.0872	0.0827	2.2444	0.1113	1.4202	0.1265	0.8432	0.0590
O-IBIS-SVD	0.0654	0.0661	0.6281	0.0674	0.5894	0.0664	0.4386	0.0646
	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	$\boldsymbol{\theta}_4$	$\boldsymbol{\theta}_5$	$\boldsymbol{\theta}_6$	$\boldsymbol{\theta}_7$	
LW	2.3992	2.2664	2.4326	1.5799	1.4677	1.0942	0.9465	
Storvik	2.2014	1.9334	1.8348	1.2845	1.1505	0.8670	0.7523	
PL	2.1434	1.9534	1.8750	1.2432	1.1344	0.8384	0.7161	

Table 16.6: Mean Monte Carlo Mean Absolute Error (MCMAE) for the parameter and state estimation with $n = 50$ runs for different particle filters with the temperature dataset A with $N_p = 5000, N_{obs} = 2304$ and $N_p = 1000$ for O-IBIS-SVD



(a) $\theta^{(1)}$ (b) $\theta^{(2)}$

(c) $\theta^{(3)}$ (d) $\theta^{(4)}$

(e) $\theta^{(5)}$ (f) $\theta^{(6)}$

(g) $\theta^{(7)}$

Figure 16.17: State component k-step ahead forecast on the temperature dataset A ($k = 804$). Colour lines are the forecast density mean, shaded areas the 90% equi-tailed credibility interval. Black line is the PMMH state posterior mean estimation.

Method	$\overline{\boldsymbol{\theta}}_1$	$\overline{\boldsymbol{\theta}}_2$	$\overline{\boldsymbol{\theta}}_3$	$\overline{\boldsymbol{\theta}}_4$
LW	28.041 (0.1829)	4.5378 (0.1502)	-3.3978 (0.2526)	-0.5499 (0.1495)
Storvik	28.3945 (0.1087)	4.358 (0.1298)	-3.1534 (0.1343)	-0.6689 (0.1201)
PL	28.4803 (0.1402)	4.2895 (0.1262)	-3.0171 (0.1803)	-0.6814 (0.103)
O-IBIS	28.3494 (0.0032)	4.4663 (0.0029)	-3.0495 (0.0097)	-0.723 (0.0046)
PMMH	27.7698	4.9824	-3.5846	-0.5912
	$\overline{\boldsymbol{\theta}}_5$	$\overline{\boldsymbol{\theta}}_6$	$\overline{\boldsymbol{\theta}}_7$	
LW	-1.2096 (0.125)	0.3624 (0.0911)	-0.4372 (0.0794)	
Storvik	-1.3028 (0.1025)	0.3089 (0.0692)	-0.3923 (0.0529)	
PL	-1.2692 (0.1127)	0.3051 (0.0634)	-0.4471 (0.0776)	
O-IBIS	-1.1668 (0.0057)	0.2977 (0.0018)	-0.5024 (0.0037)	
PMMH	-1.2435	0.2277	-0.3628	

Table 16.7: Average state posterior mean (at $t = N_{obs}$) for $n = 50$ runs of the online filters for the temperature dataset (standard error in brackets) compared to the PMMH state posterior mean estimation.

Method	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$	$\overline{\mathsf{W}}_4$
LW	0.1895 (0.0083)	0.1792 (0.0071)	4.5467 (0.5903)	0.1854 (0.0078)
Storvik	0.0803 (0.0018)	0.0796 (0.0019)	1.6331 (0.1858)	0.0795 (0.0015)
PL	0.0622 (0.0015)	0.0592 (0.0012)	2.5989 (0.6529)	0.0625 (0.0015)
O-IBIS	0.0321 (3×10^{-4})	0.0322 (3×10^{-4})	0.5668 (0.0197)	0.0318 (3×10^{-4})
PMMH	0.0155	0.015	0.4481	0.0139
	$\overline{\mathsf{W}}_5$	$\overline{\mathsf{W}}_6$	$\overline{\mathsf{W}}_7$	$\overline{\mathsf{W}}_8$
LW	2.2662 (0.1053)	0.1744 (0.0068)	1.4305 (0.0777)	0.1274 (0.0062)
Storvik	1.6059 (0.1826)	0.0803 (0.0015)	0.8981 (0.0835)	0.0508 (8×10^{-4})
PL	1.4266 (0.258)	0.0586 (0.0012)	0.8081 (0.0937)	0.0375 (7×10^{-4})
O-IBIS	0.3795 (0.0106)	0.0322 (3×10^{-4})	0.2458 (0.0053)	0.0273 (3×10^{-4})
PMMH	0.2295	0.0138	0.0774	0.0144

Table 16.8: Average parameter posterior mean (at $t = N_{obs}$) for $n = 50$ runs of the online filters for the temperature dataset (standard error in brackets) compared to the PMMH parameter posterior mean estimation.

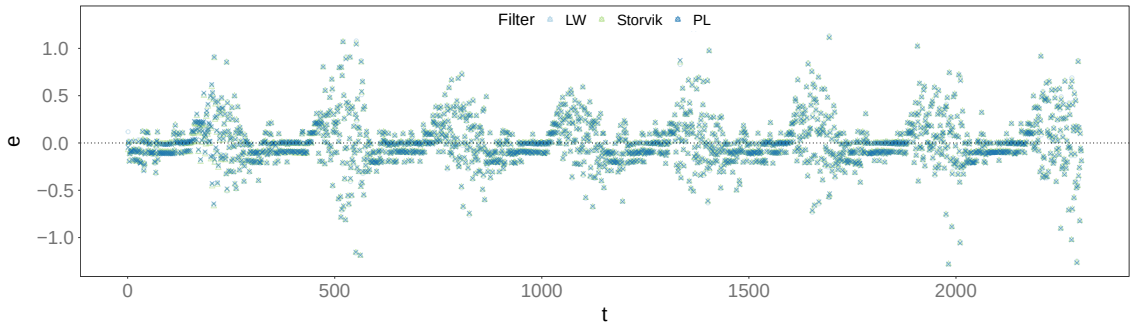
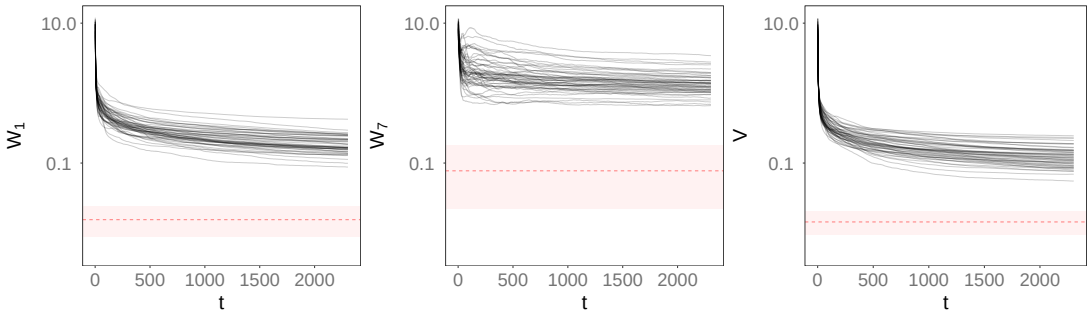


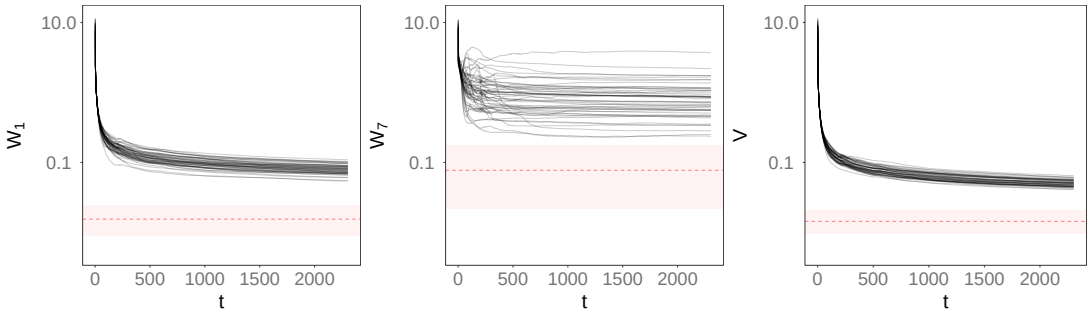
Figure 16.18: $e = (y - \hat{y})$ for one-step ahead forecast for the online filters with the temperature data

dataset A was specifically chosen due to it's regularity and absence of empirical outliers.

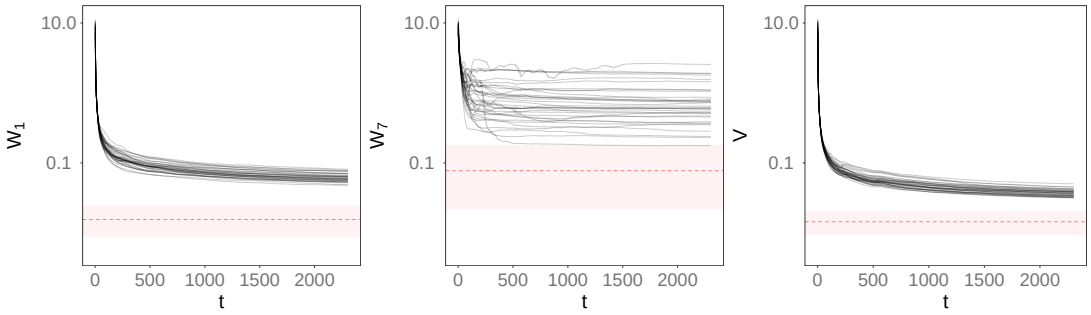
When looking at the temperature dataset A (Figure 16.7 on page 197) it is possible to see that the variability increases with the temperature value. This can potentially create problems for the state and parameter estimation by exacerbating particle impoverishment in those regions. This can verified by looking at the $\widehat{ESS}$ in Figure 16.14 on page 208, where we have a cyclical dramatic decrease of the $\widehat{ESS}$ coinciding with the high temperature values. The same behaviour can also be visually verified with other measures, such as the one-step ahead observation forecast errors in Figure 16.18 and also on the discrepancy plot in Figure 16.21 on page 217. Although, at most, the higher valued observations reach a discrepancy value $d \approx 2$, they do not reach our threshold of $d = 3$ to be considered outliers.



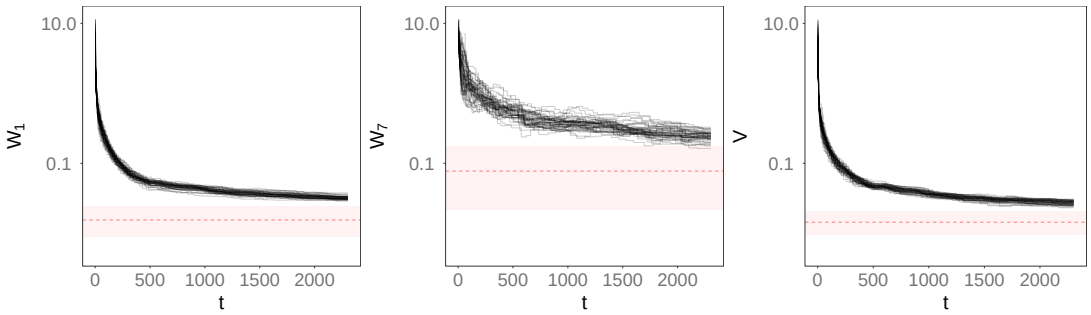
(a) LW estimation history for $\left\{\mathsf{W}^{(1)}, \mathsf{W}^{(2)}, \mathsf{W}^{(3)}\right\}$ using $n = 50$ runs



(b) Storvik estimation history for $\left\{\mathsf{W}^{(1)}, \mathsf{W}^{(2)}, \mathsf{W}^{(3)}\right\}$ using $n = 50$ runs

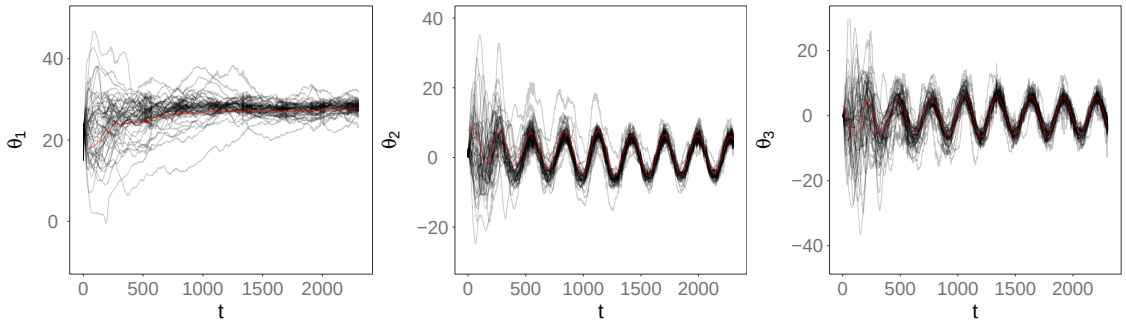


(c) PL estimation for history $\left\{\mathsf{W}^{(1)}, \mathsf{W}^{(2)}, \mathsf{W}^{(3)}\right\}$ using $n = 50$ runs

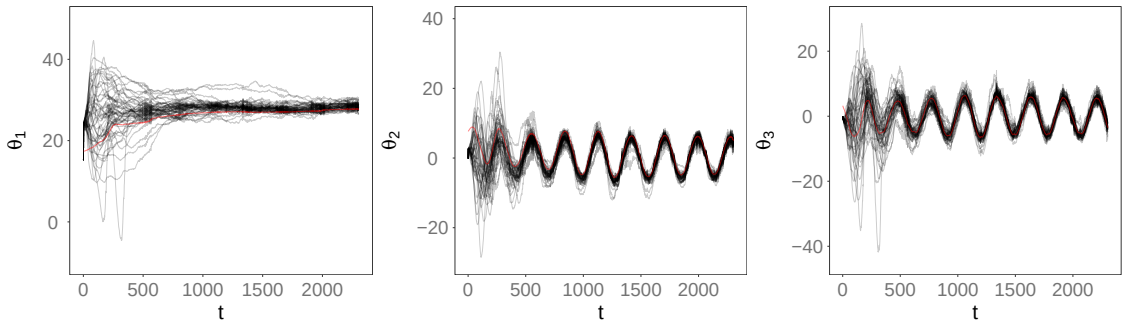


(d) O-IBIS-SVD estimation history for $\left\{\mathsf{W}^{(1)}, \mathsf{W}^{(2)}, \mathsf{W}^{(3)}\right\}$ using $n = 50$ runs

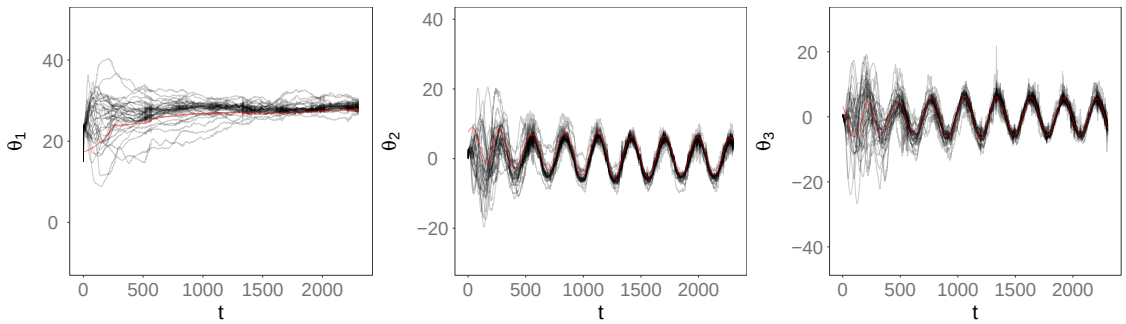
Figure 16.19: Variability in parameter posterior mean estimation history from LW, Storvik, PL and O-IBIS-SVD for $n = 50$ consecutive runs, using the temperature dataset A (*log-scale*). Red line represents PMMH posterior mean and red band the 90% equitailed credibility interval.



(a) LW estimation history for $\{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\theta}^3\}$ using $n = 50$ runs



(b) Storvik estimation history for $\{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\theta}^3\}$ using $n = 50$ runs



(c) PL estimation for history $\{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\theta}^3\}$ using $n = 50$ runs

Figure 16.20: Variability in state posterior mean estimation history from LW, Storvik and PL for $n = 50$ consecutive runs, using the temperature dataset A (*log-scale*). Red line represents the PMMH state posterior mean.

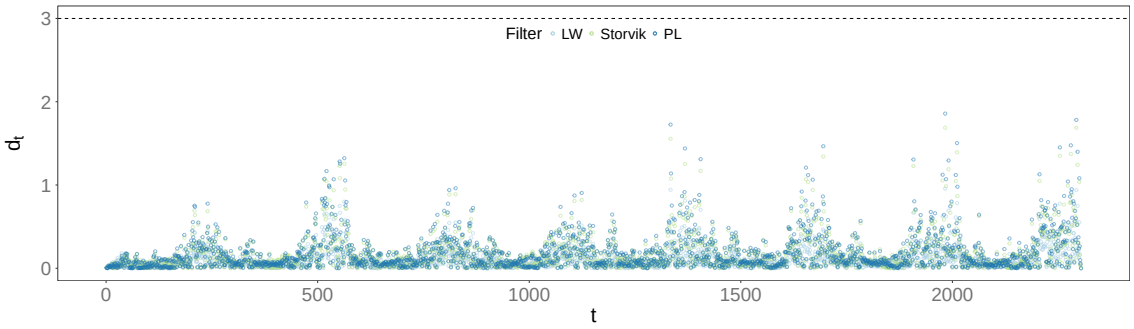


Figure 16.21: Discrepancy values for temperature dataset A, using LW, Storvik and PL. Dashed line represent anomaly threshold of $d = 3$.

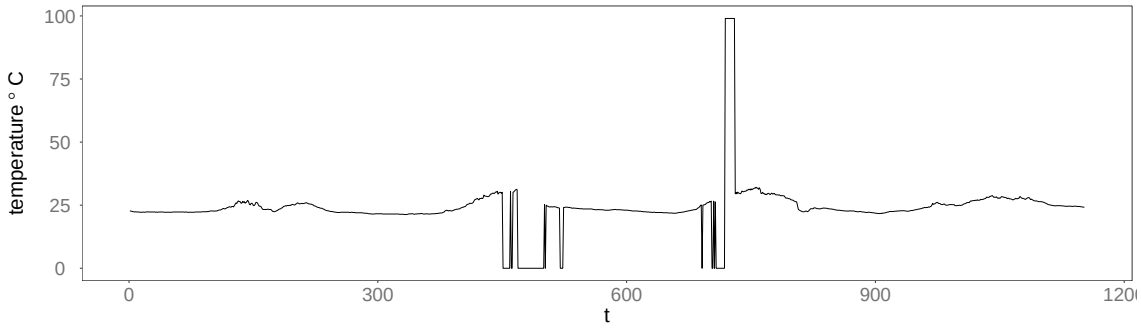


Figure 16.22: 2015 temperature data for Austin, Texas (USA) for $47870 \leq t \leq 49022$.

16.3.2 Temperature dataset B

The second analysis performed with the temperature dataset was restricted to the data between time-points $t = 47870$ and $t = 49022$. This data had obvious outliers, as visually verifiable in Figure 16.22. These values could possibly result from sensor malfunctions, since all of them have consistently the values of either 0°C or 99°C. For the subsequent analysis a simpler model was used than the one in Section 16.3.1. A Normal DLM was used, also including an underlying mean and a seasonal component, but this time consisting of a single harmonic, $h = 1$. This model can then be represented by

$$y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi} \sim \mathcal{N}\left(\mathsf{F}^T\boldsymbol{\theta}_t, V\right)$$
$$\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi} \sim \mathcal{N}\left(\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}\right)$$

with structural matrices

$$\mathsf{F} = \begin{bmatrix}1 & 1 & 0\end{bmatrix}^T, \qquad \mathsf{G} = \begin{bmatrix}1 & 0 & 0 \\ 0 & \cos\frac{2\pi}{p} & \sin\frac{2\pi}{p} \\ 0 & -\sin\frac{2\pi}{p} & \cos\frac{2\pi}{p}\end{bmatrix}, \qquad p = 288.$$

As in Section 16.3.1.1 on page 198, offline estimation was performed using PMMH. The PMMH used a bootstrap filter with $N_p = 4000$ particles, a state prior of $\boldsymbol{\theta}_0 \sim \mathcal{N}\left(\mathbf{m}_0, \mathsf{C}_0\right)$ where

$$\mathbf{m}_0 = \begin{bmatrix}15 & 0 & 0\end{bmatrix}^T, \qquad \mathsf{C}_0 = \operatorname{diag}\left(100, 100, 100\right).$$

The traces for the PMMH estimation can be found on Appendix A.5.2 on page 284. Online estimation was performed using the Liu and West, Storvik and Particle Learning filters using the same model structure as state priors. All of the online filters used $N_p = 4 \times 10^4$

with resampling at static checkpoints $n = 1$. The parameter priors for both the online and offline filters were:

$$\begin{aligned} \sigma_0^2 &\sim \mathcal{IG}(1,1) \\ \mathsf{W}_0 &\sim \mathcal{IW}(3, \mathbf{I}_3). \end{aligned}$$

As in Section 16.3.1.1 on page 198, the fully adapted version of the filters was used.

The comparison of the estimated posteriors $p(\boldsymbol{\Phi}|\mathcal{D}_T)$ between offline and online methods can be viewed in Figure 16.23 and the estimation history for the marginal $p(\boldsymbol{\Phi}|\mathcal{D}_t)$ can be viewed in Figure 16.27.

16.3.2.1 Offline estimation

Offline estimation for $\boldsymbol{\Phi} = \{\mathsf{W}, V\}$ was performed using PMMH, MCWM, IBIS, O-IBIS (and their SVD implementations) and EM. The parameters' marginal posteriors at $t = N_{obs}$ can be view in Figure 16.23 and are summarised in Table 16.9. The traces and autocorrelation plots for PMMH can be viewed in the Appendix A.5.2 on page 284.

As with temperature dataset A, IBIS and O-IBIS were both performed with a number of particles $N_p = 2 \times 10^4$ and in the O-IBIS case the observation window was $h = 500$ observations, both using state priors $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{0}, 100\boldsymbol{I})$. The resampler used for IBIS and O-IBIS was the systematic resampler and the resampling-move step occurred at $\widehat{ESS} < N_p/2$. PMMH used a bootstrap filter with $N_p = 4000$ with a systematic resampler at static checkpoint $n = 1$. The priors used with IBIS, O-IBIS, PMMH and MCWM for $\boldsymbol{\Phi}_0$ where

$$\tau_\mu^2, \tau_{d,1:7}^2 \sim \mathcal{IG}(1,1), \qquad V = \sigma^2 \sim \mathcal{IG}(1,1).$$

Also, similarly to the previous section, IBIS and O-IBIS where also implement and evaluated using the KF-SVD implementation as well as the standard KF, named respectively IBIS-SVD and O-IBIS-SVD. The EM algorithm initial parameter set was $\boldsymbol{\Phi}_0 = 100$.

We can see from Table 16.9 and Figures 16.23 that the results from PMMH and MCWM are consistent (with the exception of the higher harmonic W_3), but that the EM and IBIS/O-IBIS/SVD parameter estimation values differ greatly from the PMMH and MCWM results. An exception is the first harmonic of $\mathcal{F}(288, 1)$ where the EM value is consistent with PMMH and MCWM and the second harmonic where the O-IBIS value is also consistent with MCWM. Since the variance of the PMMH parameter estimation is high, there is a overlap in the values for $\mathsf{W}_1, \mathsf{W}_2$ and W_3 for the EM, O-IBIS and PMM-

[9] The EM method does not provide a posterior for $\boldsymbol{\Phi}$. The value corresponds to the point estimate at final iteration.

Method	Time (s)	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$	$\overline{V}$
IBIS	514.95	0.0519	0.0384	1.7079	0.0429
IBIS-SVD	1095.77	0.0519	0.0384	1.7079	0.0429
O-IBIS	629.12	0.1931	0.0472	24.716	0.0288
O-IBIS-SVD	1252.06	0.0450	0.0400	1.4465	0.0486
EM[9]	42.65	0.2293	11.7082	0.0469	4.1391
PMMH	–	12.6907	11.5926	34.8308	29.6739
MCWM	–	17.8303	11.8323	10.5831	29.3088

Table 16.9: Summary of the parameter posterior means using offline estimation methods (including O-IBIS) for temperature dataset B data.

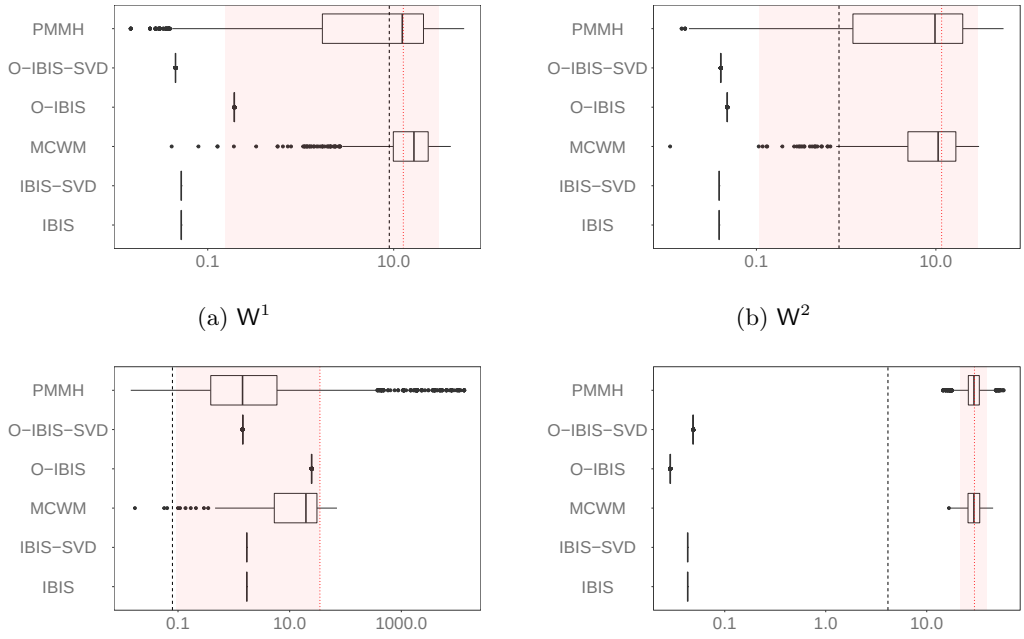
H/MCWM, although, notably not for V.

With this dataset, we can see from Figure 16.24 that is there is a clear collapse of the IBIS/SVD and O-IBIS/SVD methods when dealing with extreme outliers. Although these methods are based on resample-move scheme (thereby "rejuvenating" the parameters), the $N_{\boldsymbol{\Phi}}$ KFs used for the likelihood estimation collapse to a single value of $\left\{\mathbf{m}_t^{(i)}, \mathsf{C}_t^{(i)}\right\}$ when encountering the first outlier. The IBIS class of algorithm's parameter estimation in this case collapsed close to boundaries of the PMMH estimation for W (with should be evaluated with caution, due to the nature of the dataset), however this was not the case for V.

Regarding computational times, we can see from Table 16.9 that these values are also atypical. In this case, the "online" versions of IBIS had a higher computational cost due to the fact that after the collapse of the filter, the $\widehat{ESS}$ was never below the threshold, effectively never triggering the rejuvenation step. The EM estimation is particularly inconsistent with PMMH for V and we observe that W_2 initialises with a mean value close the PMMH final posterior estimation it stabilises in the "correct" value.

16.3.2.2 Online estimation

As with temperature dataset A, online state and parameter estimation was performed using the LW, Storvik and PL filters. The model structure is the same as used for the offline estimation and all of the filters were tested with $N_p = 4 \times 10^4$, LW shrinkage factor was $\delta = 0.99$ and stratified resampling was used with a static checkpoint of $n = 1$. The



(a) W^1 (b) W^2

(c) W^3 (d) V

Figure 16.23: Estimated parameter posteriors for $\boldsymbol{\Phi}$ at $t = N_{obs}$ for the online methods, compared to the offline methods on the temperature dataset B. Vertical black line represents the EM estimation, red line the PMMH mean and shaded area the PMMH 90% equitailed credibility interval.

parameter priors for all three methods were

$$\sigma_0^2 \sim \mathcal{IG}(1,1)$$
$$\mathsf{W}_0 \sim \mathcal{IW}(3, \mathbf{I}_3).$$

The filters used the optimal proposal density. The parameter estimation results compared to PMMH/MCWM and computational costs are presented in Table 16.11 and the state estimation MSE and computational costs when compared to PMMH/MCWM are presented in Table 16.10.

We can see from Figure 16.28 and Table 16.11 that for parameter W_1 PL and Storvik are consistent with the PMMH estimation. An interesting behaviour is noticeable in Figure 16.27 that, in the case of W_1, when the first outlier occurs, the LW filter collapses while the sufficient statistics-based method quickly converge to the PMMH posterior for W_1. A similar behaviour can be observed for W_2, while in this case the LW filter collapses within the boundaries of the PMMH estimation. For parameter W_3, again we see the collapse of LW to a single particle, while Storvik and PL avoid collapse, but still show substantial degeneracy signs. Finally, for V, we see the collapse of LW and the sufficient statistics

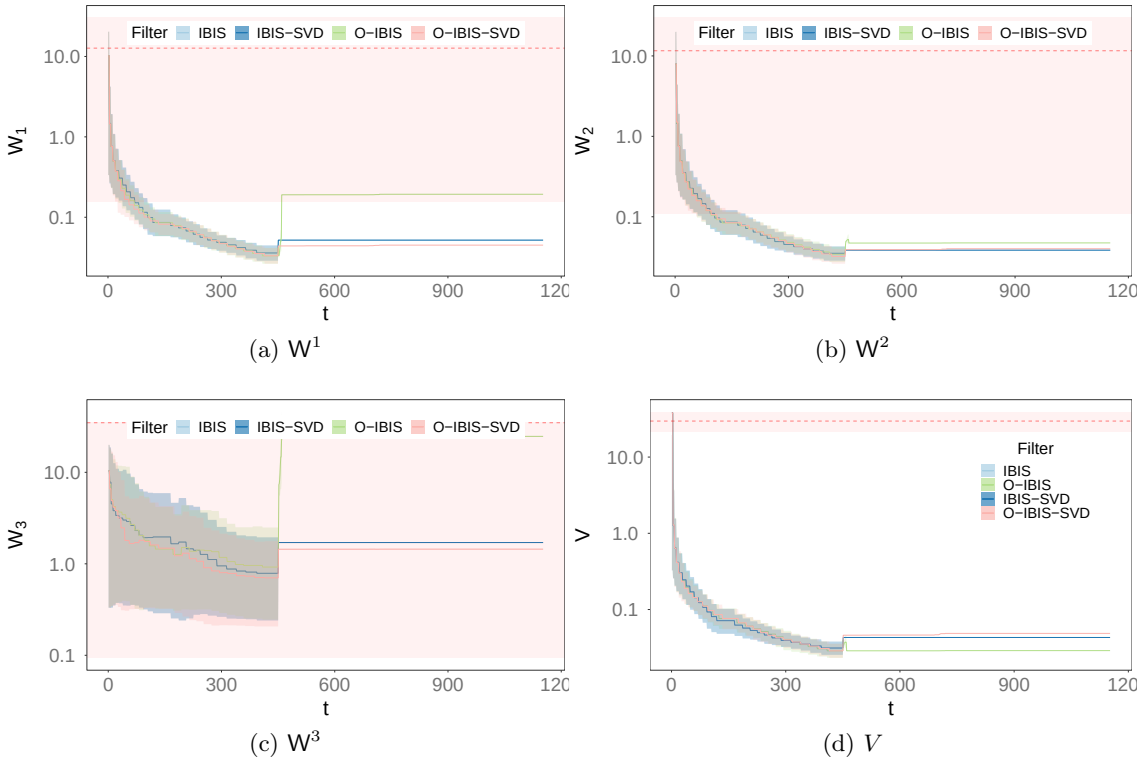


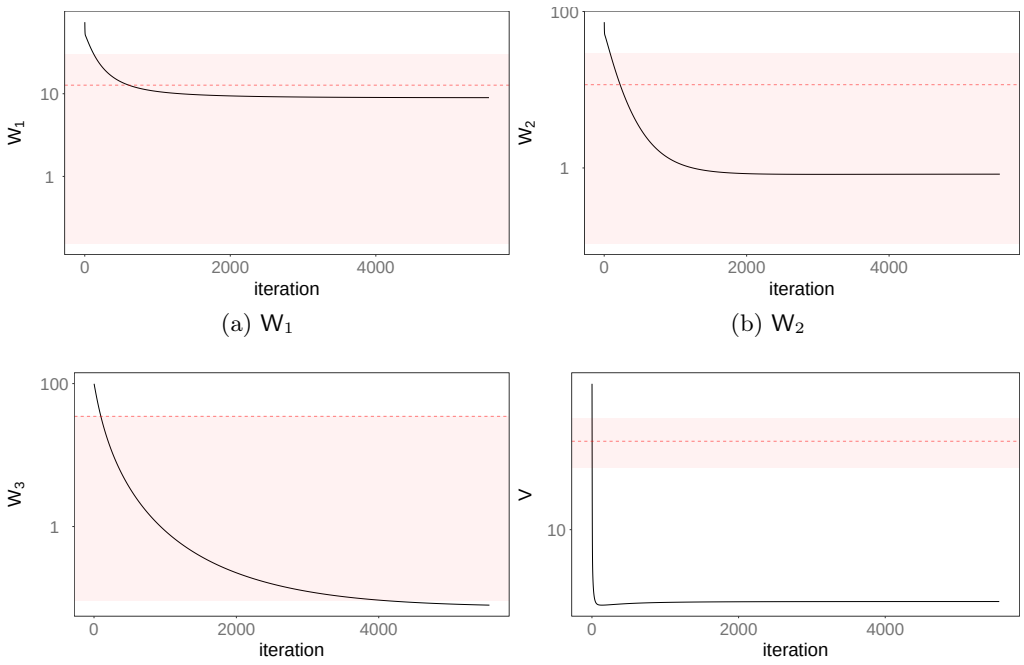
Figure 16.24: Estimation history for parameters $\boldsymbol{\Phi} = \{\mathsf{W}, V\}$ using IBIS/SVD and O-IBIS/SVD for temperature dataset B. Solid lines represent the posterior mean and shaded areas the 90% equitailed credibility interval. PMMH in red (dashed horizontal line is the posterior mean and shaded area the 90% quantile interval).

methods converging to the PMMH value but stabilising in a substantially different value.

Regarding the computational cost of the three filters, we see similar results, as expected with the analysis for the temperature dataset A, with LW having a slight advantage and PL and Storvik showing similar execution times (with PL slightly slower than Storvik). For a model dimensionally smaller than temperature dataset A, we see that the cost per iteration is between 116.46 to 331.17 milliseconds, still within the criteria for near real-time state and parameter estimation.

Regarding state estimation, we can verify (in Table 16.10) that although collapse occurs for the all filters, LW and Storvik are generally better in terms of state estimation accuracy, while from the sufficient statistics-based methods, Storvik shows higher accuracy than PL when compared to the PMMH state estimation. From Figure 16.26 we can see that after final set of outliers PL filtering density estimation diverge strongly from the PMMH estimation, while LW and Storvik approximate better the PMMH estimation. A consequence of this is the PL filter's high MSE value in Table 16.10.

Regarding state forecast, all forecasts are performed after the outliers have occurred.



(a) W_1 (b) W_2

(c) W_3 (d) V

Figure 16.25: Estimation history for parameters $\boldsymbol{\Phi} = \{\mathsf{W}, V\}$ using EM for the temperature dataset B. Horizontal dashed line represents the PMMH posterior mean and shaded area the 90% equitailed credibility interval.

We see from Figure 16.29 that although the estimation for $\boldsymbol{\theta}_1$ is consistent with PMMH, this is not the case for seasonal components $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$.

16.3.2.3 Forecast

Regarding the k-step ahead forecast a $k = 252$ step state forecast was performed and can be seen in Figure 16.29. The effect of particle collapse is clearly visible by the manifest mismatch between the forecast values and the PMMH state estimates, specially for $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$.

16.3.2.4 Discrepancies

To detect potential anomalies online, the discrepancy value at each time t, $d(y_t)$, was calculated at each iteration according to the Normal DLM (as detailed in Section 2.4.3.2). Anomalies where flagged when the discrepancy value was $d(y_t) > 3$ and the results for the LW, Storvik and PL filters can be viewed respectively in Figures 16.31, 16.32 and 16.33. The vertical grey lines represent the empirical outliers, classified whenever $y_t = 0$ or $y_t = 99$.

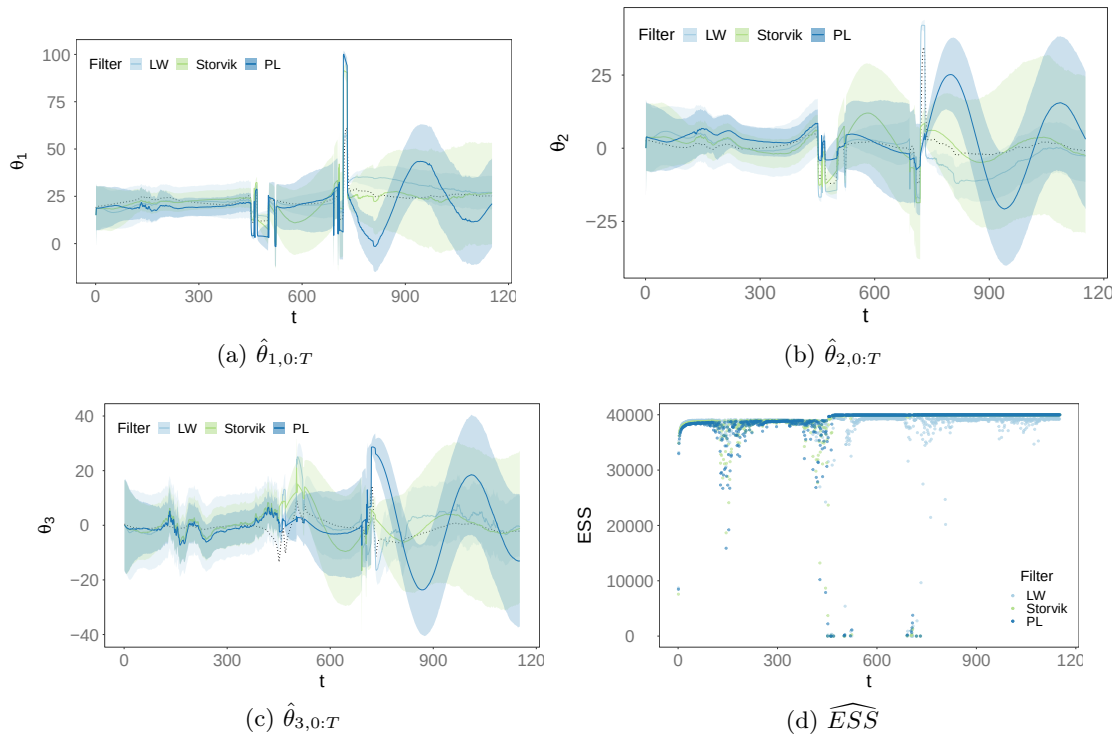
Method	**MSE**			**Time**	
	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	*iteration (ms)*	*total (s)*
LW	18.9086	22.474	23.6664	116.46	134.169
Storvik	31.1403	16.2002	23.272	326.84	376.525
PL	110.9315	90.9919	96.5557	331.17	381.519

Table 16.10: MSE for different particle filters with the temperature dataset B compared to the PMMH estimation and computational time for $N_p = 4 \times 10^4, N_{obs} = 1152$.

Method	**Time (s)**	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$	$\overline{V}$
LW	134.169	0.1569	0.2367	2.2001	0.0985
Storvik	376.525	16.2631	1.0984	1.5765	0.352
PL	381.519	19.6032	0.5532	0.9106	0.2006
PMMH	–	12.6907	11.5926	34.8308	29.6739
MCWM	–	17.830	11.832	10.583	29.308

Table 16.11: Parameter posterior mean estimation and computation time with online methods for the temperature data B.

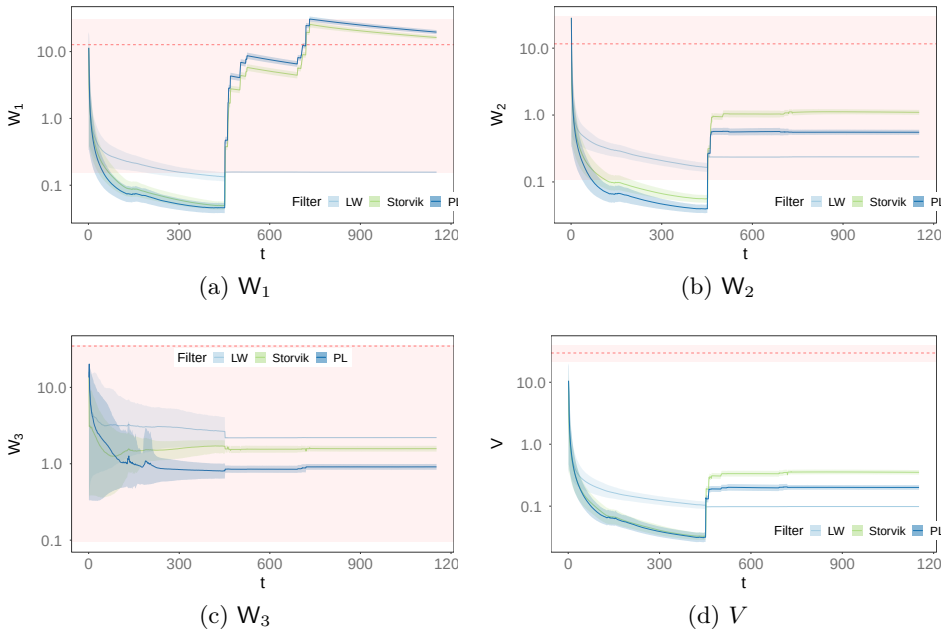
From the discrepancy data, we can see that LW flagged most of the empirical outliers correctly, while Storvik and PL showed consistent results between them. Storvik and PL also display a considerable number of "borderline" outliers, when using the above criteria with $d(y_t)$ close, but below 3.



(a) $\hat{\theta}_{1,0:T}$ (b) $\hat{\theta}_{2,0:T}$

(c) $\hat{\theta}_{3,0:T}$ (d) $\widehat{ESS}$

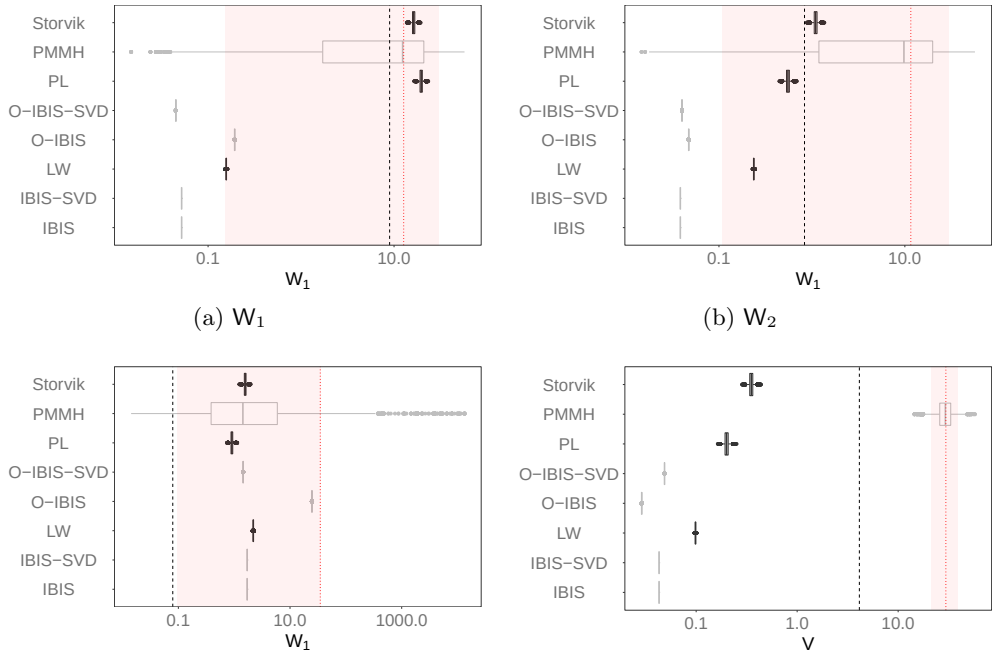
Figure 16.26: State components for PF estimation on the temperature dataset B ($N_{obs} = 1152$) and $\widehat{ESS}$. Colour lines represent state posterior mean and shaded areas 90% equitailed credibility interval. Dashed line represents PMMH state posterior mean



(a) W_1 (b) W_2

(c) W_3 (d) V

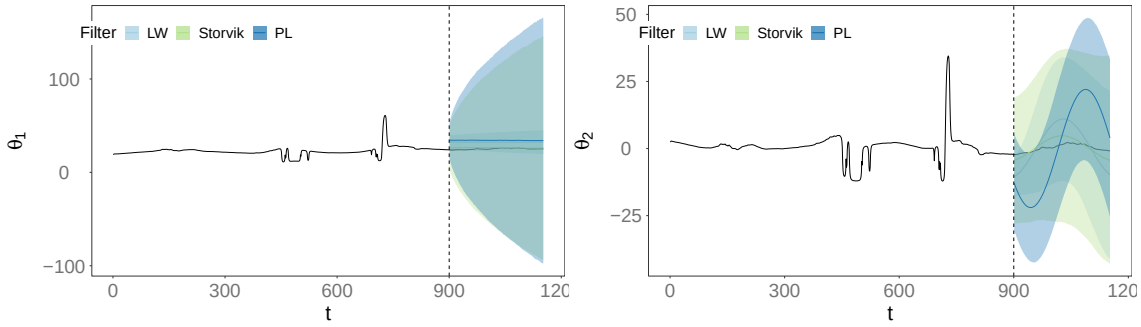
Figure 16.27: Estimation history for $\boldsymbol{\Phi}$ using the online methods for the temperature dataset B. Horizontal line represents the PMMH and red band the 90% equitailed credibility interval. Colour lines represent the parameter posterior mean estimation and shaded areas the 90% quantile interval.



(a) W_1 (b) W_2

(c) W_3 (d) V

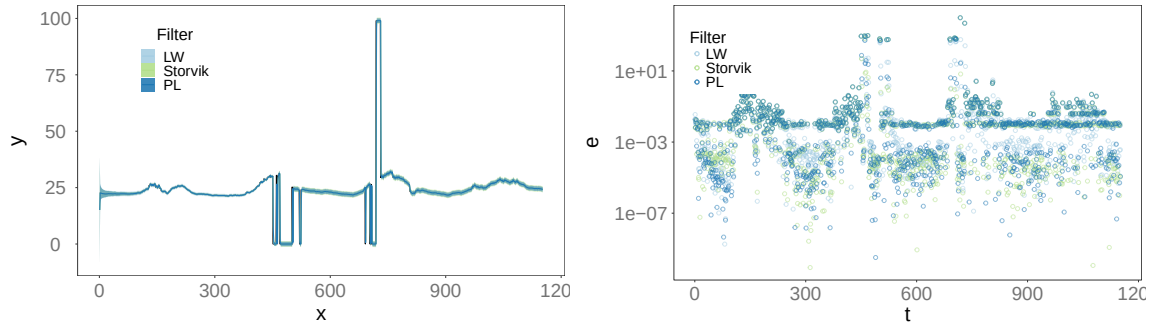
Figure 16.28: Estimated marginals for $\boldsymbol{\Phi}$ at $t = N_{obs}$ for the online methods compared to the offline methods on the temperature dataset B. Vertical black line represents the EM estimation, red dashed line the PMMH mean and vertical red band the PMMH posterior 90% equitailed credibility interval.



(a) $\boldsymbol{\theta}_1$ (b) $\boldsymbol{\theta}_2$

(c) $\boldsymbol{\theta}_3$ (d) y_{t+k}

Figure 16.29: State component k-step ahead forecast on the temperature dataset B. Solid black line is the PMMH state posterior mean, colour lines represent the forecast density mean and shaded area the forecast 90% equitailed credibility interval.



(a) $\hat{y}_{t+1}$ (b) $e_t = (y_t - \hat{y}_t)$

Figure 16.30: One-step ahead forecast value (*left*) and respective errors (*right*) for the online filters with the temperature dataset B.

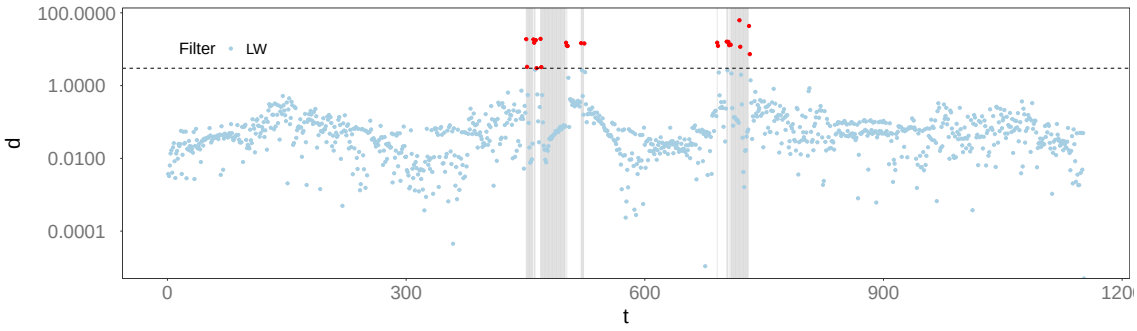


Figure 16.31: Discrepancy values ($d(y_t) > 3$ threshold) for temperature dataset B using the LW filter

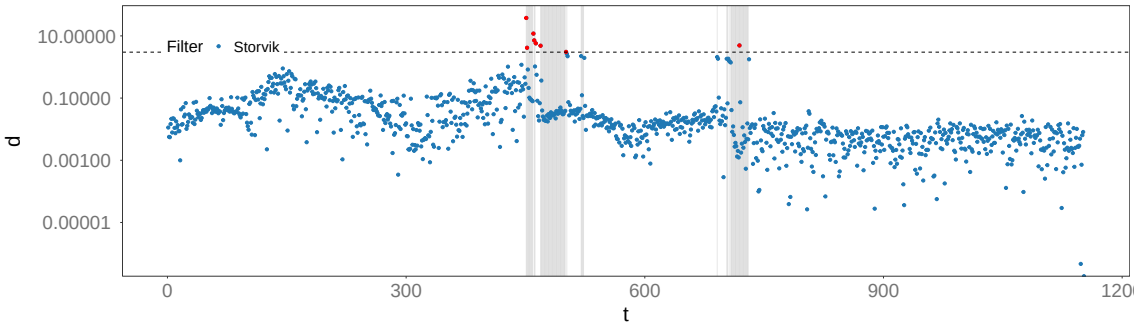


Figure 16.32: Discrepancy values ($d(y_t) > 3$ threshold) for temperature dataset B using the Storvik filter

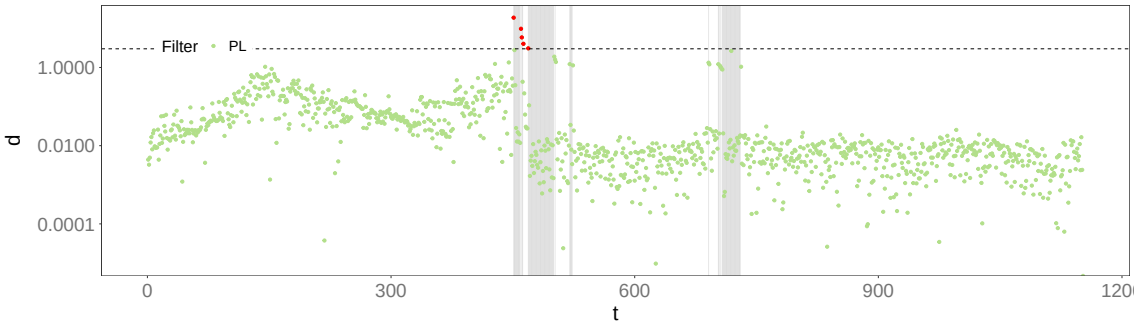


Figure 16.33: Discrepancy values ($d(y_t) > 3$ threshold) for temperature dataset B using the PL filter

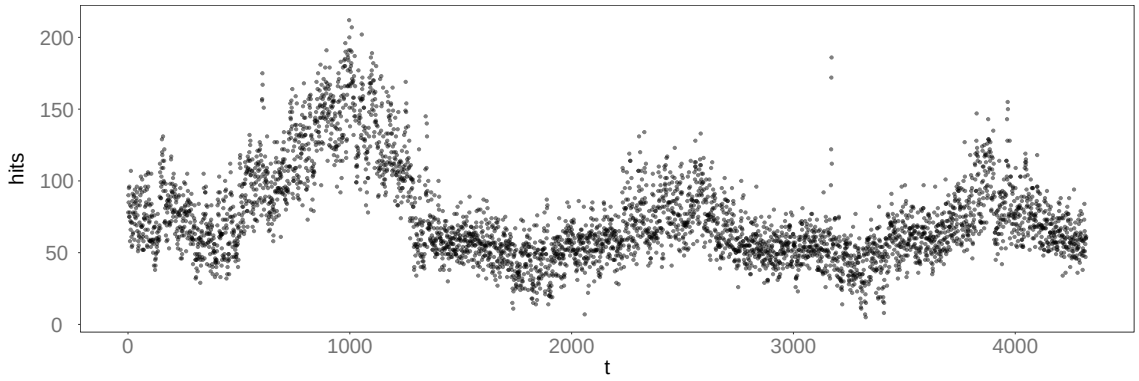


Figure 16.34: World Cup 1998 dataset

16.4 World Cup 1998

To test the estimation methods using the Poisson DLM, we have used in this section the World Cup 1998[10] (WC98) dataset (Arlitt & Jin (2000)). The data contains records of Hyper Text Transfer Protocol (HTTP) access requests to the official web server for the 1998 football World Cup. The data comprises access from April 30, 1998 and July 26, 1998 totalling 1,352,804,107 requests.

The original data captured visits to web resources as event type data but was converted to a time series with granularity of one minute. Resources other than web pages (*e.g.* images) were removed and the remaining resources aggregated in that time window. The resulting data is shown in Figure 16.34.

The model chosen was a Poisson DLM following the relations of Section 2.3.2 on page 17, that is

$$
\begin{aligned}
y_t|\boldsymbol{\theta}_t, \boldsymbol{\Phi} &\sim \mathrm{Po}\left(e^{\eta_t}\right) \\
\eta_t &= \mathsf{F}^T\boldsymbol{\theta}_t \\
\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}\right)
\end{aligned}
$$

Since the data subset analysed corresponds to approximately 3 days we use a locally constant component to capture the underlying mean along with a Fourier component to capture the daily effect. As the data is sampled every minute, the period of the Fourier components will be $p = 1440$, representing the daily pattern. As in Section Section 16.3.1 on page 196, we have measured[11] the forecast performance of the Storvik filter for different values of h and have chosen a number of harmonics ($h = 3$) that represents a compromise between a flexible approximation of the seasonal component and computational costs.

[10] Available at: http://ita.ee.lbl.gov/html/contrib/WorldCup.html (Accessed 13^{th} September 2017)
[11] Full results of the forecast runs available in Appendix C.2.

Therefore, according to (2.15), the model structure will be

$$\mathsf{F} = [1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0]^T,$$

$$\mathsf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \cos\frac{2\pi}{p} & \sin\frac{2\pi}{p} & 0 & 0 & 0 & 0 \\ 0 & -\sin\frac{2\pi}{p} & \cos\frac{2\pi}{p} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cos\frac{4\pi}{p} & \sin\frac{4\pi}{p} & 0 & 0 \\ 0 & 0 & 0 & -\sin\frac{4\pi}{p} & \cos\frac{4\pi}{p} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cos\frac{6\pi}{p} & \sin\frac{6\pi}{p} \\ 0 & 0 & 0 & 0 & 0 & -\sin\frac{6\pi}{p} & \cos\frac{6\pi}{p} \end{bmatrix},$$

$p = 1440$.

16.4.1 Offline estimation

Offline estimation of the parameter set $\boldsymbol{\Phi} = \{\mathsf{W}\}$ was performed using PMMH, SMC^2 and O-SMC^2. The marginals using PMMH, SMC^2 and O-SMC^2 at time $t = N_{obs}$ are presented in Figure 16.35 on page 233. The estimation history for SMC^2 and O-SMC^2 is presented respectively in Figure 16.36 on page 234. The traces and auto-correlation plots for PMMH can be viewed in Appendix A.6.

SMC^2 and O-SMC^2 were both performed with a number of particles $N_{\boldsymbol{\Phi}} = 2000$ and $N_{\boldsymbol{\theta}} = 1000$ and in the O-SMC^2 case the observation window was $h = 500$ observations. The number of particles for O/SMC^2 were chosen based on keeping the computational cost of the non-rejuvenating steps still comparable to the online methods. The priors used with SMC^2, O-SMC^2 and PMMH for $\boldsymbol{\Phi}_0$ were

$$\mathsf{W}_{0,1:7} \sim \mathcal{IG}(1,1).$$

For the $N_{\boldsymbol{\Phi}}$ particles filters of SMC^2 and O-SMC^2 we have used a state prior

$$\begin{aligned} \boldsymbol{\theta}_0 &\sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0), \\ \mathbf{m_0} &= [10 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T \qquad (16.4) \\ \mathsf{C}_0 &= \mathrm{diag}(100, 10, 10, 10, 10, 10, 10) \qquad (16.5) \end{aligned}$$

with the $\mathsf{C}_{0,1,1} = 100$ to indicate a potential higher variability of the underlying mean when

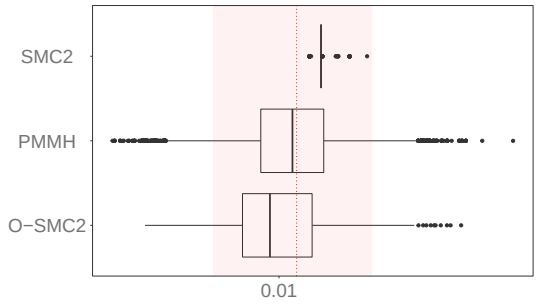
Method	**Time (s)**	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$	$\overline{\mathsf{W}}_4$
SMC^2	84888.98	0.0117 (0.0002)	0.0135 (0.0001)	0.4297 (0.015)	0.0123 (0.0001)
O-SMC^2	42457.1	0.0089 (0.002)	0.0091 (0.0015)	0.2123 (0.1063)	0.0098 (0.0019)
PMMH	–	0.0107 (0.002)	0.0107 (0.002)	0.3247 (0.442)	0.0106 (0.002)
Method	**Time (s)**	$\overline{\mathsf{W}}_5$	$\overline{\mathsf{W}}_6$	$\overline{\mathsf{W}}_7$	
SMC^2	84888.98	0.1160 (0.0087)	0.0117 (0.0002)	0.0949 (0.01)	
O-SMC^2	42457.1	0.0601 (0.0409)	0.0065 (0.0016)	0.0519 (0.0299)	
PMMH	–	0.1897 (0.1418)	0.0106 (0.002)	0.0986 (0.0644)	

Table 16.12: Parameter posterior mean estimation (at $t = N_{obs}$) with IBIS and O-IBIS and posterior mean with PMMH (standard deviation in brackets) for the WC98 data.

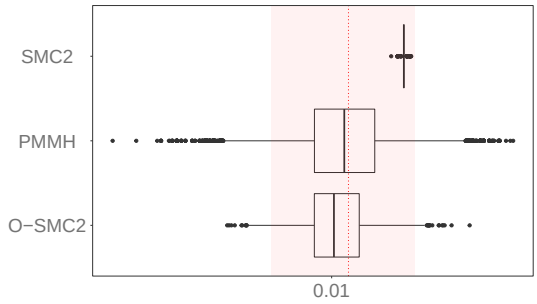
compared to the seasonality components. A summary of parameter estimation results for the offline methods is presented in Table 16.12.

We see in Table 16.12 the considerable difference in computational cost between SMC^2 and O-SMC^2. In terms of the parameter estimation itself, we can see that the results of SMC^2 and O-SMC^2 are generally comparable. Both methods provide a reasonable approximation of the PMMH estimated posteriors as we can see from Figure 16.36. For all the separate components of W, we can also see from Figure 16.35 that there is a considerable overlap between the SMC^2, O-SMC^2 and PMMH estimates. However, SMC^2 suffers from a severe underestimation of the parameter posterior variance when compared to PMMH. This could be due to an insufficient number of particles and precisely a scenario where the automatic calibration of $N_{\boldsymbol{\theta}}$, as described in Chapter 15, could be potentially useful. It is also noteworthy that both methods use the same N_{Φ} and N_{θ}. However, when the rejuvenation stage occurs at a some time t_r, O-SMC^2 is estimating the likelihood over h data points while SMC^2 is estimating it over t_r, potentially with $t_r > h$. The situation might occur[12] where N_{θ} is adequate for a window of size h but not t_r. In this scenario, the insufficient N_{θ} for SMC^2 can lead to a high estimator variance and a low acceptance rate, which can contribute to the observed posterior variance underestimation.

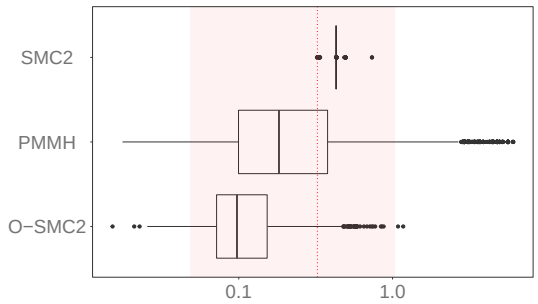
[12]In our case the last rejuvenating step of SMC^2 occurred at $t = 4282$ and O-SMC^2 at $t = 4119$, consequently SMC^2 performed the rejuvenation over 4282 observations and O-SMC^2 over 500.



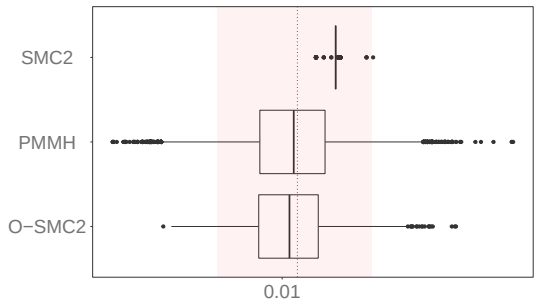
(a) W_1



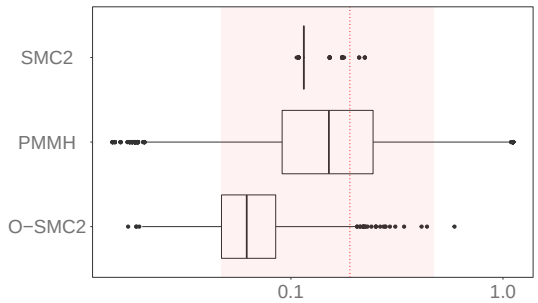
(b) W_2



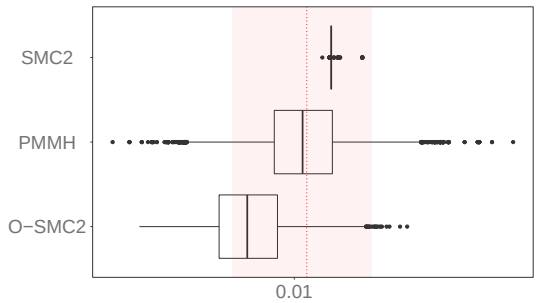
(c) W_3



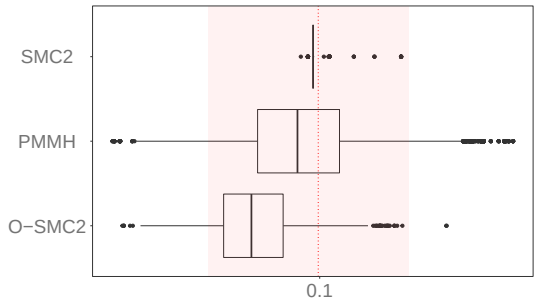
(d) W_4



(e) W_5

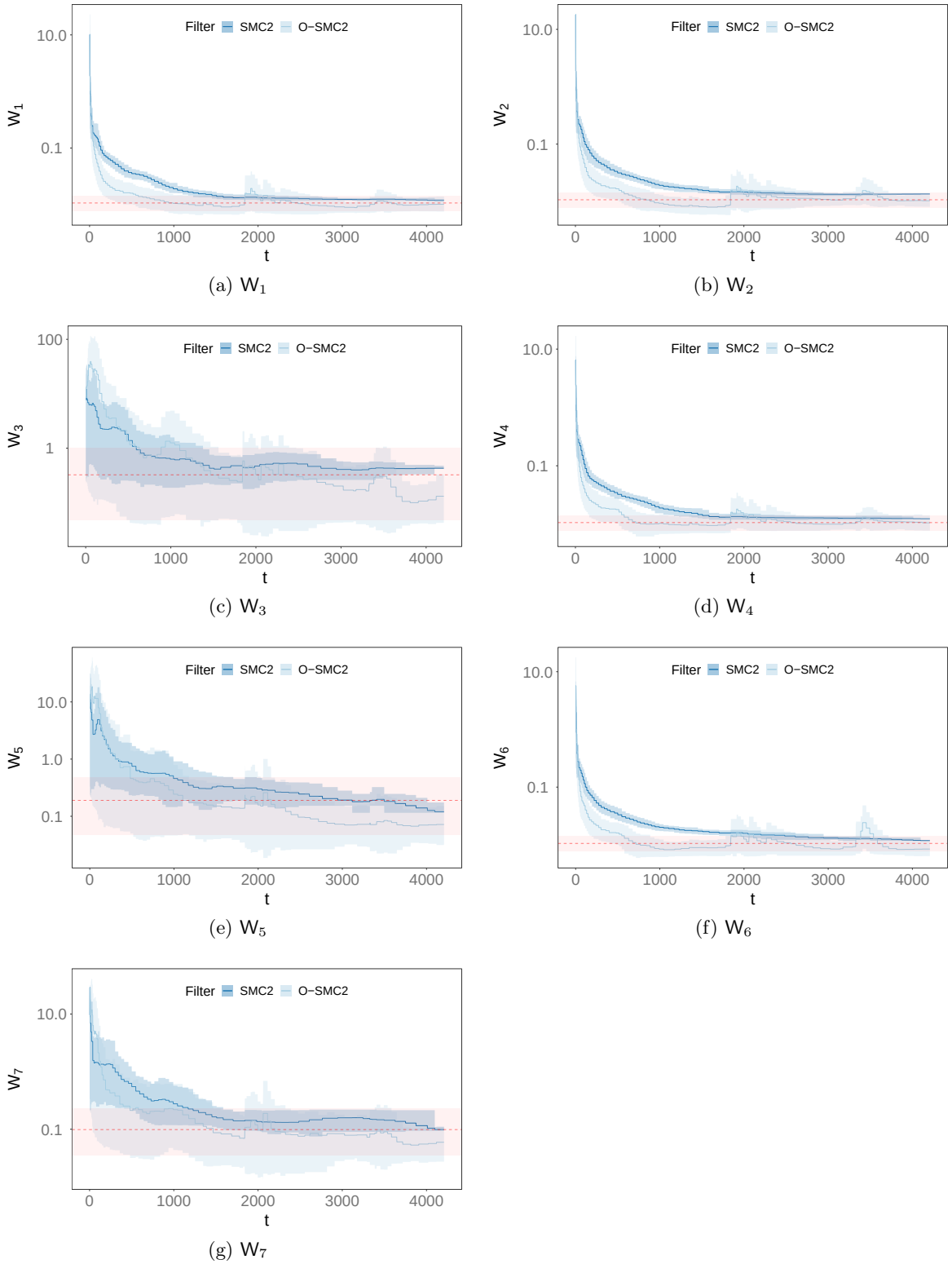


(f) W_6



(g) W_7

Figure 16.35: SMC^2 and O-SMC^2 $\boldsymbol{\Phi} = \{\mathsf{W}\}$ parameter posterior estimation at $t = N_{obs}$ and PMMH parameter posterior for the WC98 dataset. Vertical red line represents the PMMH posterior mean and red band the 90% equitailed credibility interval..



(a) W_1 (b) W_2

(c) W_3 (d) W_4

(e) W_5 (f) W_6

(g) W_7

Figure 16.36: $\boldsymbol{\Phi} = \{W\}$ parameter posterior estimation history using SMC^2 and O-SMC^2 for the WC98 dataset. Solid lines represent the posterior mean and shaded areas the 90% equitailed credibility interval. PMMH in red (dashed horizontal line is the posterior mean and shaded area the 90% equitailed credibility interval).

16.4.2 Online estimation

Online state and parameter estimation was performed using the Liu & West, Storvik and Particle Learning filters as in Section 16.3.1.2 on page 203. The Storvik filter was implemented according to the algorithm described in Algorithm 9.1 on page 119, but in three variants. One using the prior, $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$, as the importance density, a second implementation using the CF as the importance density and finally using the EKF adjusted model prior as described in Section 6.2.5 on page 80. These implementations were named Storvik, Storvik-CF and Storvik-EKF respectively.

The model structure used was the same as the one presented in Section 16.4 on page 230. The state priors were the same for all the filters and drawn from a vague prior $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0)$ with the $\mathbf{m}_0$ and C_0 taking the same values as (16.4) and (16.5) respectively.

Several runs for varying N_p were performed[13], similarly to Section 16.3.1.2 on page 203, after which the value of $N_p = 4.5 \times 10^4$ was chosen. After performing[14] several runs of Liu & West with a range of smoothing parameters a value of $\delta = 0.99$ was chosen and the resampler used was the stratified resampler presented in Section 7.3.2 with a static checkpoint of $n = 1$, that is, performing resampling at every time step t. The parameter priors used were

$$\mathsf{W}_0 \sim \mathcal{IW}(7, 0.2\mathbf{I}_7).$$

We will present the estimation for approximately 3 days worth of minutely data ($N_{obs} = 4320$).

The filter's results are compared to a long run of a PMMH (the traces and ACF can be viewed in Appendix A.6) and the MSE is calculated considering the PMMH's result as the "true" value. In Figure 16.37 on page 237 we present the estimated marginals for the parameter set $\boldsymbol{\Phi} = \{\mathsf{W}\}$ using the online methods, in comparison to the offline results obtained in Section 16.4.1. In Figure 16.38 on page 238 we show the estimation history for $\boldsymbol{\Phi}$ using the online methods. A summary of the online parameter estimation results is presented in Table 16.13 on the next page.

The results for online state estimation are presented separately for each state component $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_7\}$ in Figure 16.38 on page 238, along with the ESS (Figure 16.41) for each of the online methods. A summary of the MSE between each state component and the PMMH state estimation is presented in Table 16.14 on page 241, along with each filter's total and iteration execution times.

We can see from Table 16.13 that PL is generally closer to the PMMH estimation (with the exception of W_7). Regarding the Storvik class, the implementation using the EKF pro-

[13] A summary of the tuning results can be found in Appendix D.2.

[14] Summary tables with the tuning results available in Appendix E.2 on page 313.

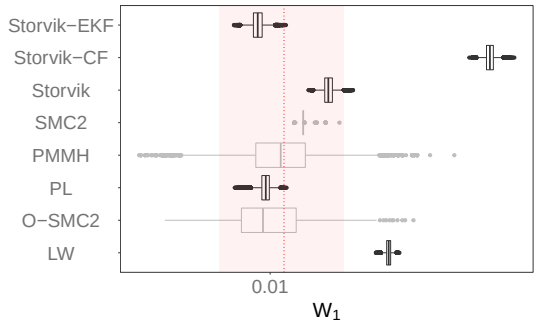
Method	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$	$\overline{\mathsf{W}}_4$	$\overline{\mathsf{W}}_5$	$\overline{\mathsf{W}}_6$	$\overline{\mathsf{W}}_7$
LW	0.018 (0.0003)	0.007 (0.0003)	1.7558 (0.3502)	0.0183 (0.0002)	0.1756 (0.0191)	0.0137 (0.0003)	0.7206 (0.1112)
Storvik	0.0134 (0.0004)	0.0097 (0.0003)	0.0866 (0.003)	0.0117 (0.0004)	0.048 (0.0013)	0.0108 (0.0003)	0.1309 (0.0043)
Storvik-CF	0.0296 (0.0008)	0.0148 (0.0004)	0.0264 (0.0012)	0.0246 (0.0008)	1.7875 (0.0642)	0.0096 (0.0003)	0.0963 (0.0042)
Storvik-EKF	0.0094 (0.0003)	0.0121 (0.0004)	0.0305 (0.0015)	0.0094 (0.0005)	0.0949 (0.0032)	0.0108 (0.0004)	0.057 (0.004)
PL	0.0098 (0.0003)	0.0104 (0.0003)	0.4679 (0.0152)	0.0147 (0.0004)	0.1382 (0.0047)	0.009 (0.0002)	0.2008 (0.0049)
PMMH	0.0107 (0.002)	0.0107 (0.002)	0.3247 (0.442)	0.0106 (0.002)	0.1897 (0.1418)	0.0106 (0.002)	0.0986 (0.0644)

Table 16.13: Summary of parameter posterior mean estimation at $t = N_{obs}$ with online estimation methods for the WC98 dataset.

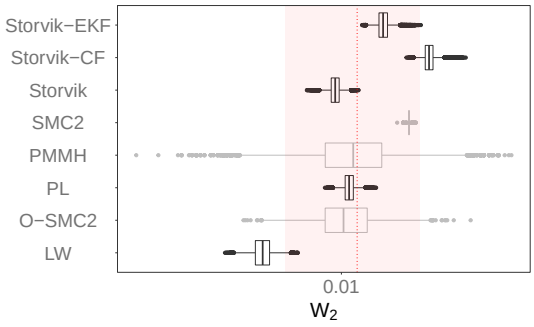
posal outperforms Storvik-CF, especially regarding the highly divergent estimate for W_5. In several W components Storvik-EKF outperforms PL, specifically in the estimation of $\mathsf{W}_3, \mathsf{W}_4$ and W_6. Regarding computational costs, we see from Table 16.14 a similar relation as with the NDLM versions, with LW having the lowest computational cost followed by Storvik and PL. Storvik-CF and Storvik-EKF do incur in the highest cost for the online methods, which is to be expected due to the additional Kalman-style recursions (and associated matrix operations) needed to create the new proposal at each time step. Still, we can see that the computational cost per iteration (*i.e.*, the average computational cost for each time step t) seems perfectly adequate to perform near real-time state and parameter estimations being, in this case, bound between 189.9 and 1057.3 milliseconds per iteration.

Regarding state estimation, we can see from Table 16.14 that when comparing the MSE between the estimated states and PMMH estimation, the sufficient-statistics-bases filters generally outperform the remaining methods. Storvik outperforms Liu and West and has a MSE closer generally closer to the PL values than Liu & West. It is noteworthy, that Storvik-EKF generally outperformed Storvik in the state estimation as it was the case with the parameter estimation. Storvik-CF did not perform better than PL (with the exception of the estimation for $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$), but in the interest of comparing different proposals within the same algorithm it did not outperform Storvik as can be seen from Table 16.14. We can see the comparison between Storvik/CF/EKF in Figure 16.40, where is it clear (especially in the case of $\boldsymbol{\theta}_1$) that Storvik-EKF's $p\left(\boldsymbol{\theta}_t|\mathcal{D}_t\right)$ estimation is closer to the PMMH estimation, although displaying a higher posterior variance than Storvik. Although Storvik-EKF outperforms PL in several components of the parameter estimation, this is not the case for the state estimation. It does however clearly produces better state estimates than Storvik.

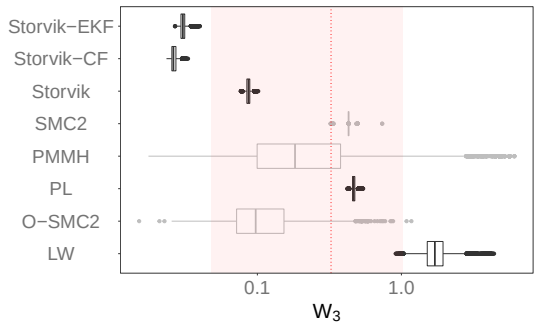
[15] Average time of non-resampling steps with $t > k$.
[16] Average time of non-resampling steps with $t > k$.



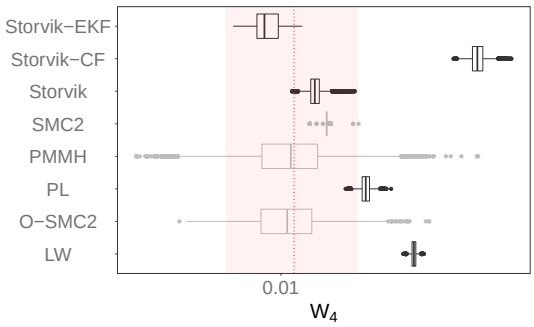
(a) W_1



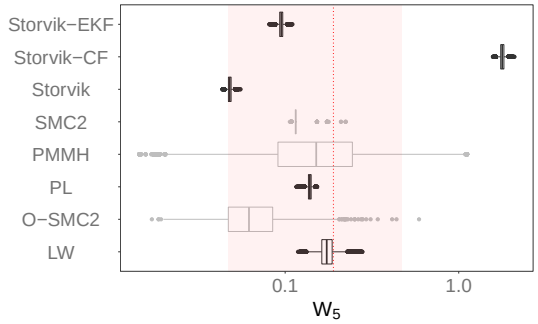
(b) W_2



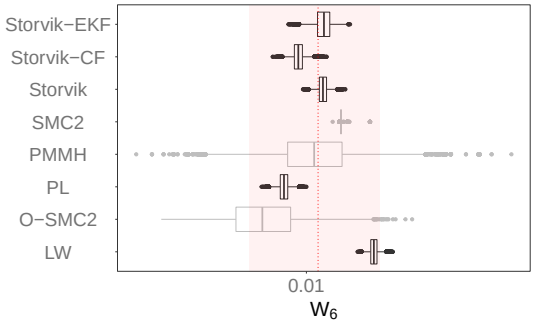
(c) W_3



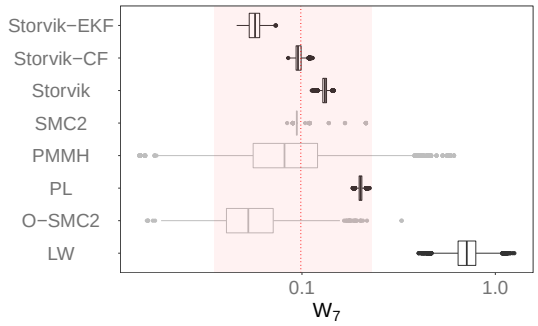
(d) W_4



(e) W_5

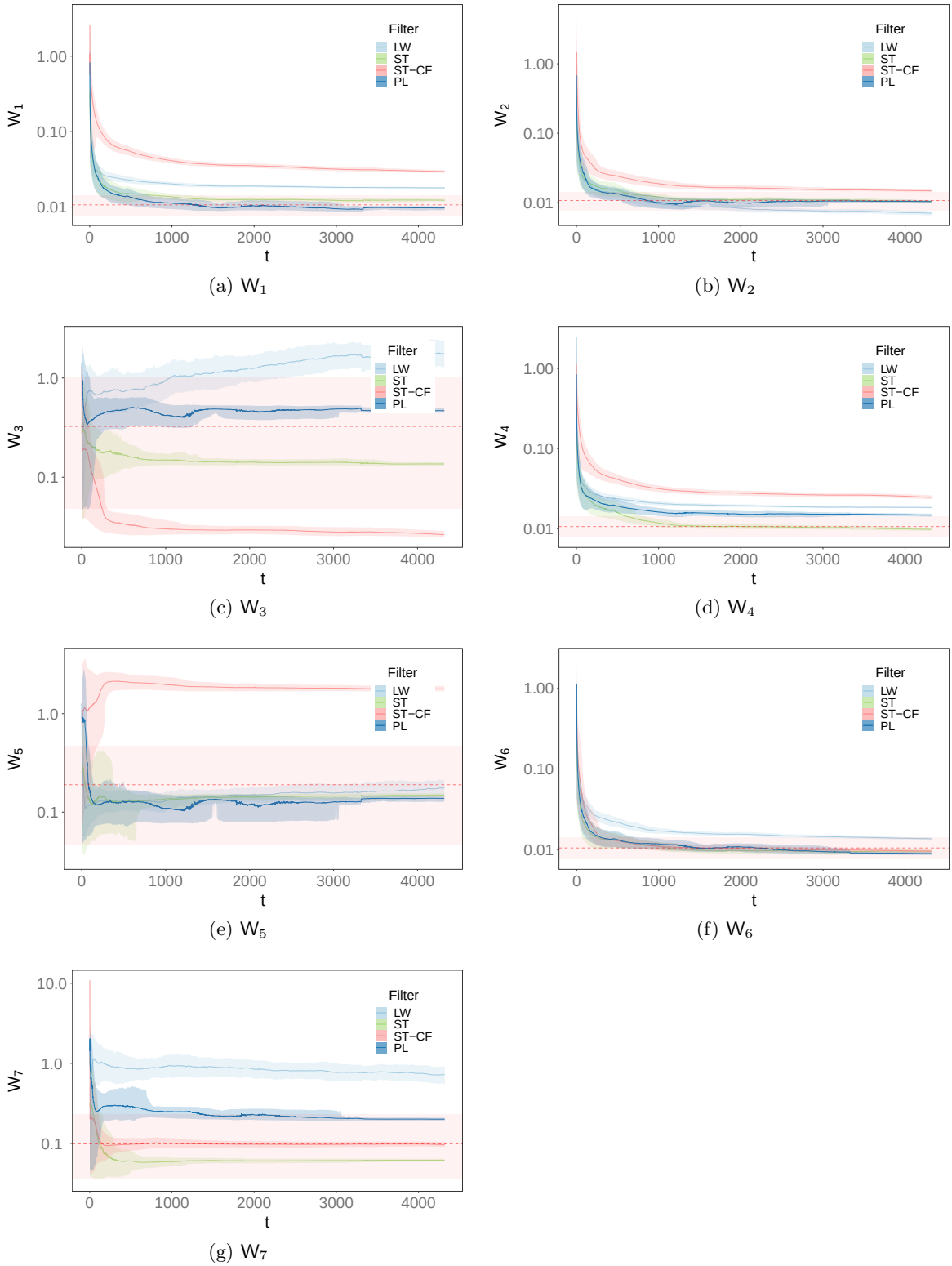


(f) W_6



(g) W_7

Figure 16.37: Parameter posterior estimation for $\boldsymbol{\Phi}$ at $t = N_{obs}$ for the online methods, compared to the offline methods with the WC98 data. Red dashed line represents PMMH parameter posterior mean and shaded area the 90% equitailed credibility interval.



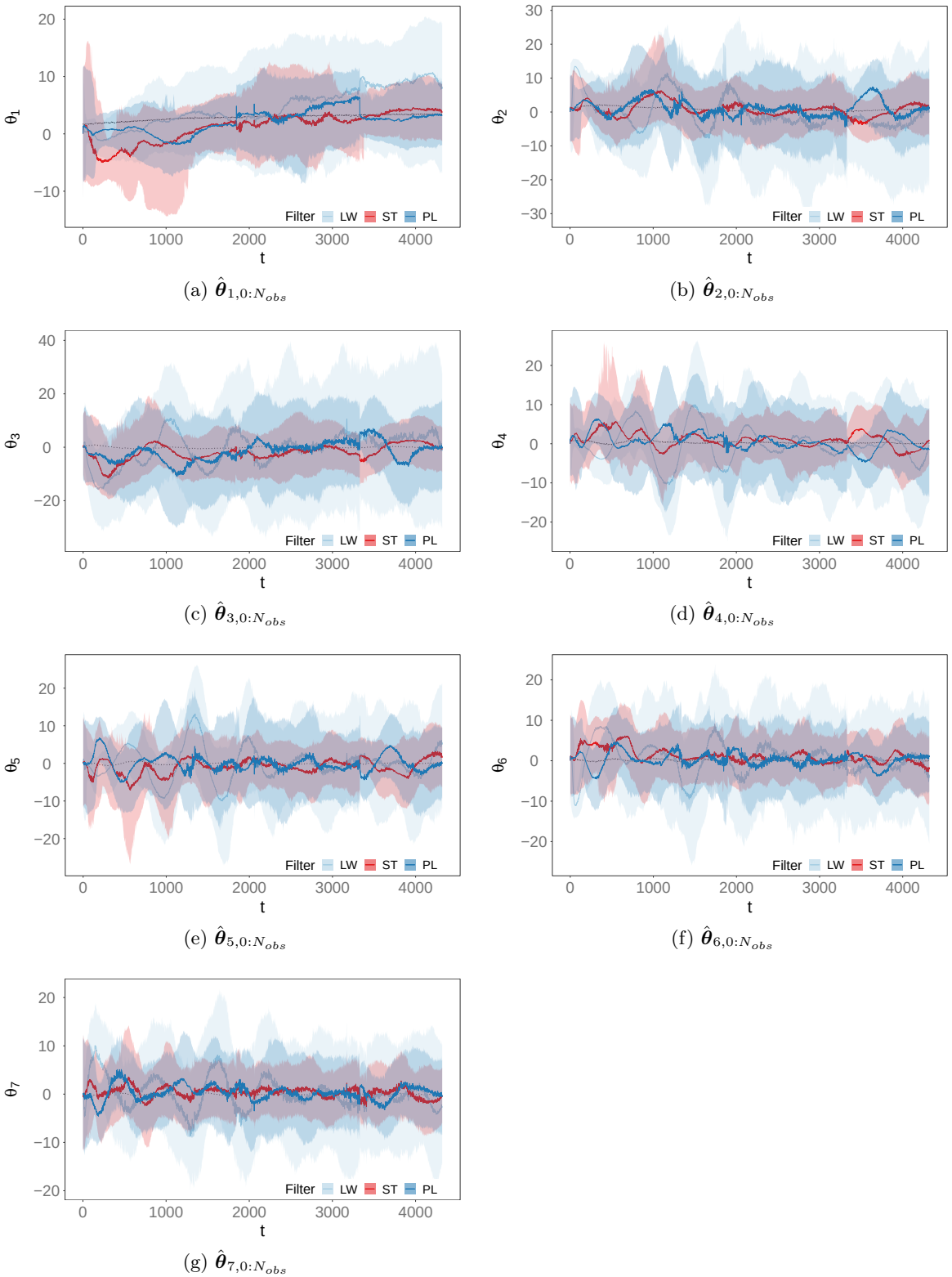
(a) W_1 (b) W_2

(c) W_3 (d) W_4

(e) W_5 (f) W_6

(g) W_7

Figure 16.38: $\boldsymbol{\Phi}$ parameter posterior estimation history using the online methods for the WC98 data. Colour lines represent the posterior mean and shaded areas the 90% equitailed credibility interval. Horizontal dashed line represent the PMMH posterior mean and shaded area the 90% equitailed credibility interval.



(a) $\hat{\boldsymbol{\theta}}_{1,0:N_{obs}}$

(b) $\hat{\boldsymbol{\theta}}_{2,0:N_{obs}}$

(c) $\hat{\boldsymbol{\theta}}_{3,0:N_{obs}}$

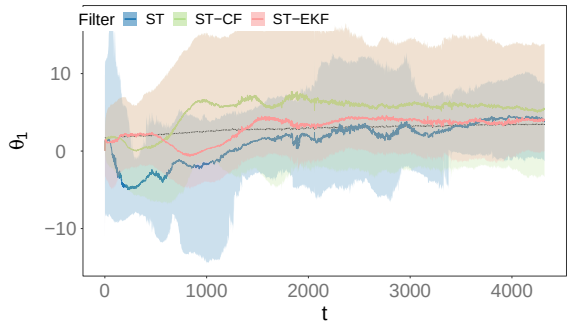
(d) $\hat{\boldsymbol{\theta}}_{4,0:N_{obs}}$

(e) $\hat{\boldsymbol{\theta}}_{5,0:N_{obs}}$

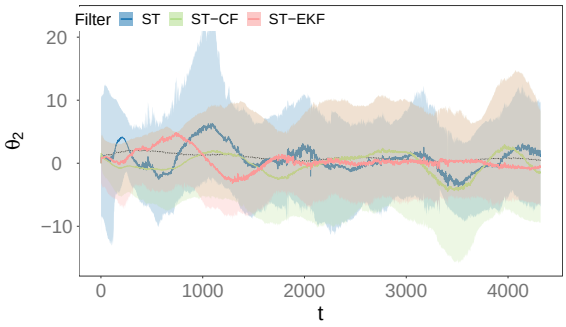
(f) $\hat{\boldsymbol{\theta}}_{6,0:N_{obs}}$

(g) $\hat{\boldsymbol{\theta}}_{7,0:N_{obs}}$

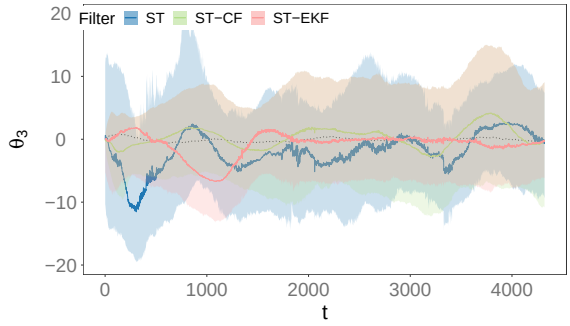
Figure 16.39: State posterior estimation history using online methods on the WC98 dataset. Colour lines represent state posterior mean and shaded areas 90% equitailed credibility interval. Dashed black line represents PMMH state posterior mean.



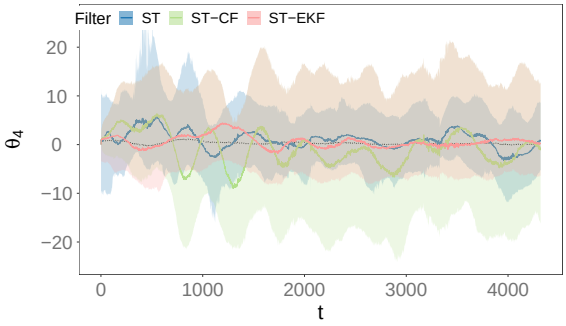
(a) $\hat{\boldsymbol{\theta}}_{1,0:N_{obs}}$



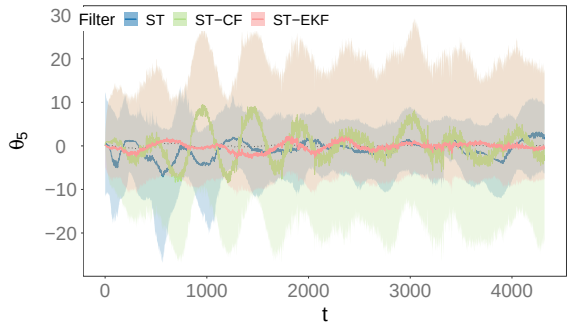
(b) $\hat{\boldsymbol{\theta}}_{2,0:N_{obs}}$



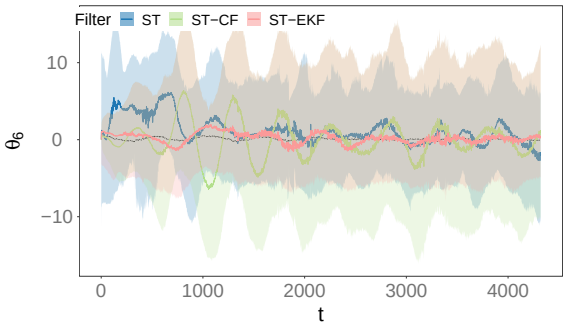
(c) $\hat{\boldsymbol{\theta}}_{3,0:N_{obs}}$



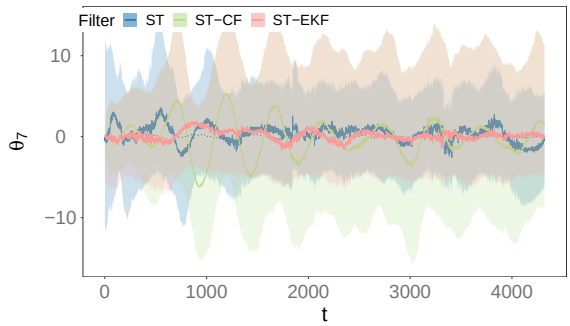
(d) $\hat{\boldsymbol{\theta}}_{4,0:N_{obs}}$



(e) $\hat{\boldsymbol{\theta}}_{5,0:N_{obs}}$



(f) $\hat{\boldsymbol{\theta}}_{6,0:N_{obs}}$

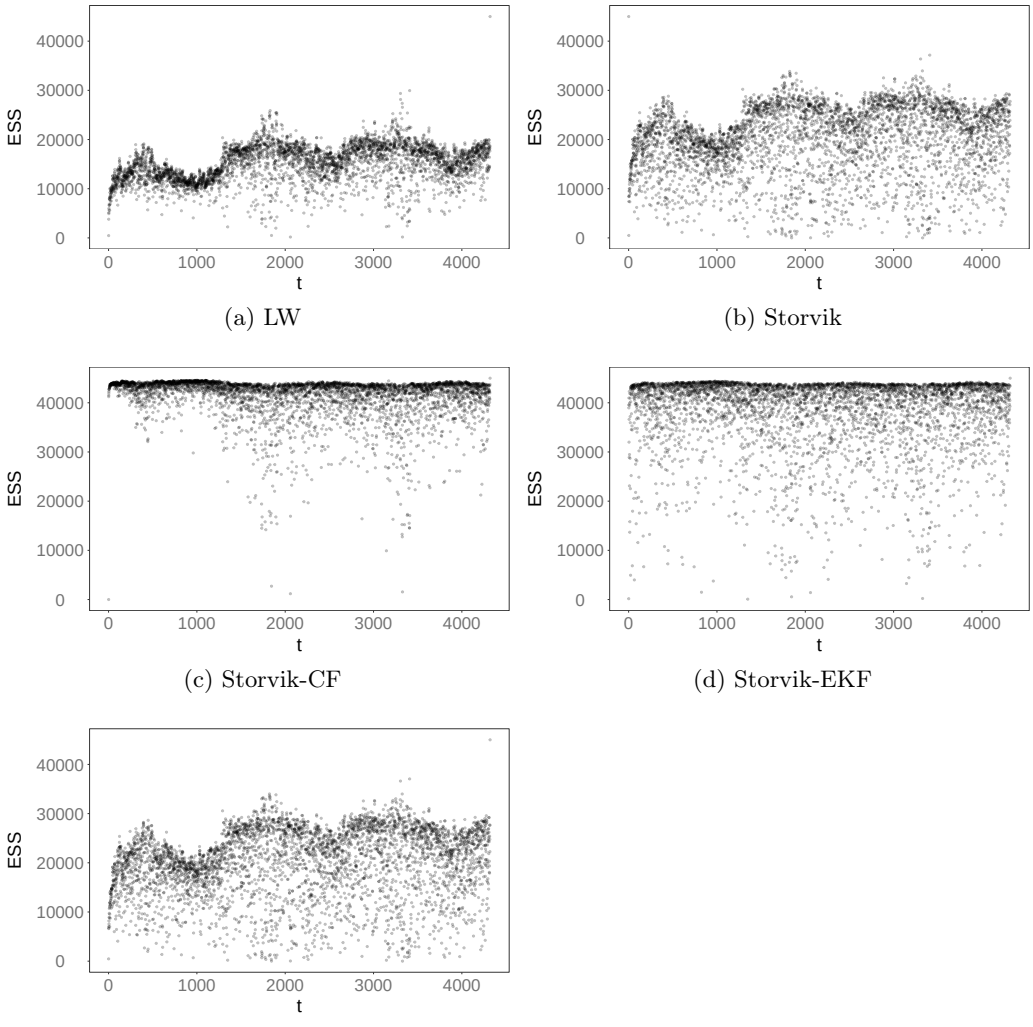


(g) $\hat{\boldsymbol{\theta}}_{7,0:N_{obs}}$

Figure 16.40: State posterior estimation history using Storvik, Storvik-CF and Storvik-EKF on the WC98 dataset. Colour lines represent state posterior mean and shaded areas 90% equitailed credibility interval. Dashed black line represents PMMH state posterior mean.

Filter	MSE							Time	
	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	$\boldsymbol{\theta}_4$	$\boldsymbol{\theta}_5$	$\boldsymbol{\theta}_6$	$\boldsymbol{\theta}_7$	*iteration (ms)*	*total (s)*
LW	10.9658	19.9896	35.7292	17.5371	18.2089	13.9044	9.7558	189.9	820.397
Storvik	7.0778	3.6249	11.6331	3.6656	4.3315	3.8827	1.2036	664.545	2870.838
Storvik-CF	7.7535	3.57	2.4655	12.5878	13.3732	5.865	5.2507	1057.3	4567.651
Storvik-EKF	1.1855	2.0548	4.2009	1.3258	0.706	0.409	0.3236	767.9	3317.423
PL	3.4092	5.021	14.2514	4.808	4.4804	2.6072	2.1838	477.3	2061.966
SMC^2	9.6067	16.7394	20.7526	7.7167	10.9001	6.2586	3.8609	4147.2^{15}	70082.64
O-SMC^2	5.6993	2.5284	0.8852	0.5111	0.2231	0.1715	0.1119	3860.9^{16}	78128.32

Table 16.14: State posterior mean MSE (relatively to PMMH state posterior mean) and computational time for different filters with the WC98 dataset.



(a) LW

(b) Storvik

(c) Storvik-CF

(d) Storvik-EKF

(e) PL

Figure 16.41: $\widehat{ESS}$ for the online methods with the WC98 data.

Method	MSE						
	$\hat{\boldsymbol{\theta}}_{1,t+k}$	$\hat{\boldsymbol{\theta}}_{2,t+k}$	$\hat{\boldsymbol{\theta}}_{3,t+k}$	$\hat{\boldsymbol{\theta}}_{4,t+k}$	$\hat{\boldsymbol{\theta}}_{5,t+k}$	$\hat{\boldsymbol{\theta}}_{6,t+k}$	$\hat{\boldsymbol{\theta}}_{7,t+k}$
LW	10.154	11.291	11.658	5.0332	4.9828	1.1868	1.1452
Storvik	0.28814	0.78964	0.38509	0.42907	0.39201	0.88338	0.90245
Storvik-CF	6.5922	1.9969	1.6356	20.381	21.181	3.8636	3.9776
Storvik-EKF	0.7255	0.3864	0.034385	0.17426	0.12706	0.10575	0.093408
PL	3.4421	2.8063	2.7061	0.69519	0.6394	0.17886	0.17393

Table 16.15: State posterior mean k-step ($k = 804$) ahead forecast Mean Squared Error (MSE) relatively to PMMH state posterior estimation for different particle filters with the WC98 dataset.

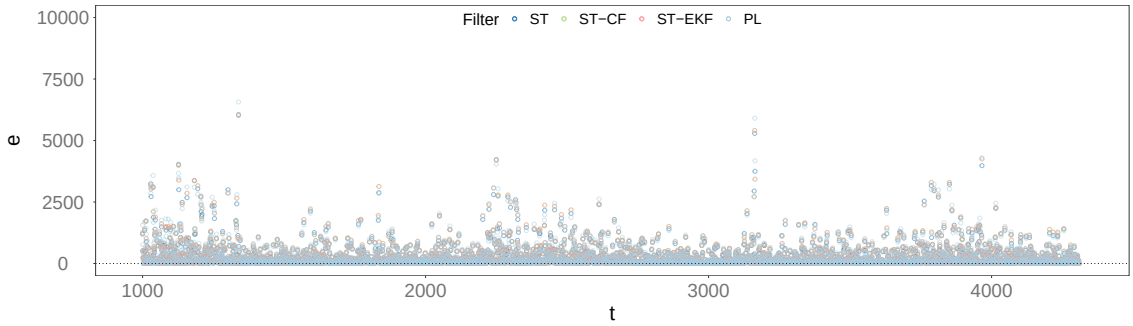
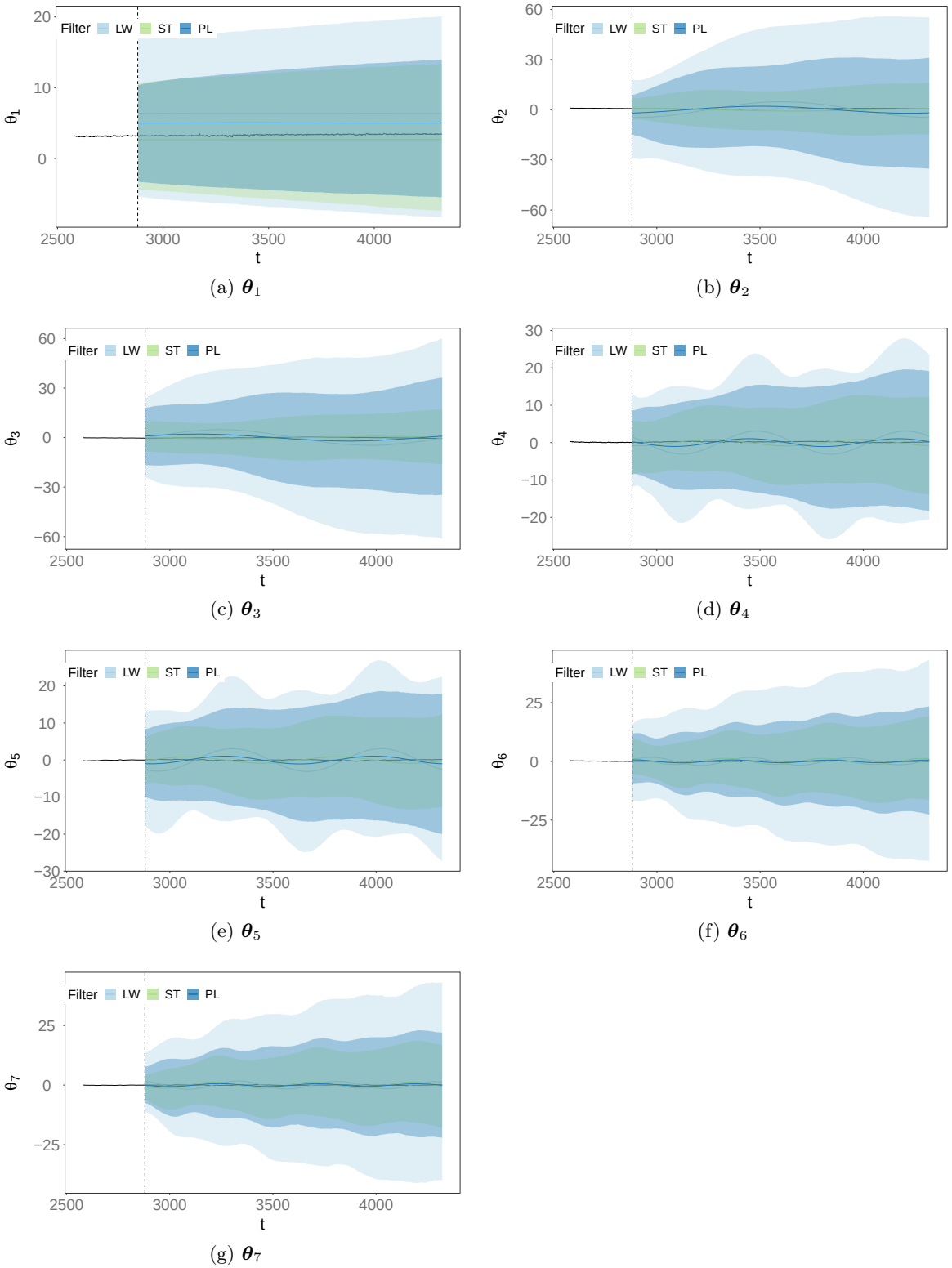


Figure 16.42: One-step ahead forecast errors for the online filters with the WC98 data.

16.4.3 Forecast

The results for the one-step ahead observation forecast errors are presented in Figure 16.42 and summarised in Table 16.15. As in the case of temperature data, a longer k-step ahead forecast ($k = 804$) was performed on this data. The results for the k-step ahead state forecast are summarised in Table 16.15.

Regarding state forecast, we can see from Table 16.15 mixed results where Storvik-EKF outperforms the remaining methods for $\{\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_7\}$ and Storvik outperforms for the state components $\boldsymbol{\theta}_1$. From Figures 16.43 and 16.44 we can see that the state forecast is consistent with the PMMH state estimate, although LW shows a higher posterior variance than the sufficient-statistics based methods. From Figure 16.44 we can see that Storvik-CF also display a much higher variance than Storvik. It is also clear from the results that Storvik-EKF outperforms Storvik in several components of state and parameter estimation, but also in the case of the state forecast, being in fact the best performing of all the methods.



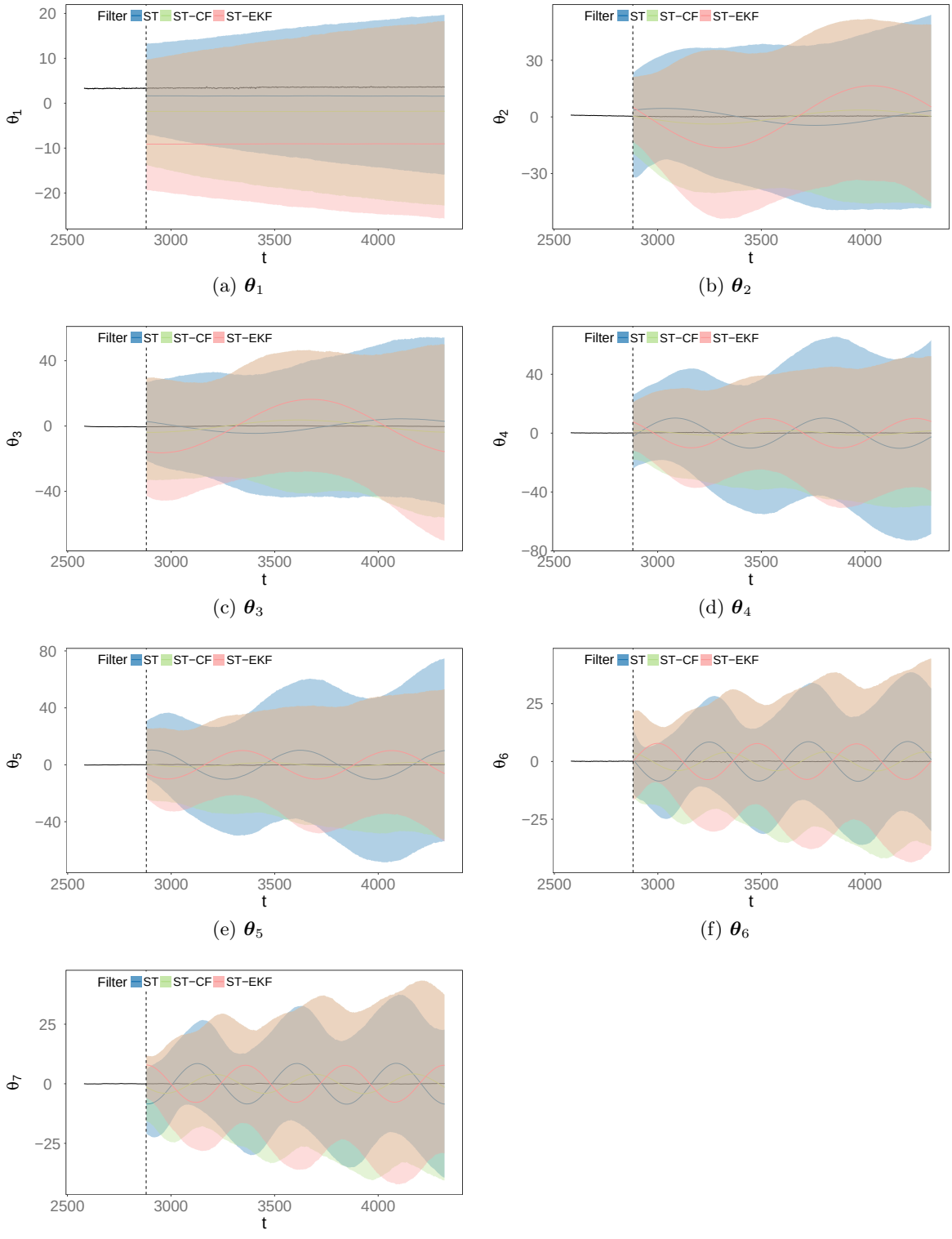
(a) $\boldsymbol{\theta}_1$ (b) $\boldsymbol{\theta}_2$

(c) $\boldsymbol{\theta}_3$ (d) $\boldsymbol{\theta}_4$

(e) $\boldsymbol{\theta}_5$ (f) $\boldsymbol{\theta}_6$

(g) $\boldsymbol{\theta}_7$

Figure 16.43: State component k-step ahead forecast on the WC98 dataset ($k = 804$). Colour lines represent posterior mean and shaded areas 90% equitailed credibility interval. Solid black line represents PMMH state posterior mean.



(a) $\boldsymbol{\theta}_1$ (b) $\boldsymbol{\theta}_2$

(c) $\boldsymbol{\theta}_3$ (d) $\boldsymbol{\theta}_4$

(e) $\boldsymbol{\theta}_5$ (f) $\boldsymbol{\theta}_6$

(g) $\boldsymbol{\theta}_7$

Figure 16.44: State component k-step ahead forecast on the WC98 dataset using Storvik and Storvik-CF ($k = 804$). Colour lines represent posterior mean and shaded areas 90% equitailed credibility interval. Solid black line represents PMMH state posterior mean.

Filter	$\overline{MCMAE}$						
	W_1	W_2	W_3	W_4	W_5	W_6	W_7
LW	0.4541	0.4327	1.5677	0.4599	2.5417	0.4423	2.1443
Storvik	0.1287	0.1226	1.4820	0.1248	1.6673	0.1276	1.4770
Storvik-CF	0.1159	0.1130	7.3560	0.1289	1.9560	0.1087	1.8907
Storvik-EKF	0.0590	0.0614	1.4177	0.0600	1.4926	0.0599	1.3529
PL	0.0865	0.0876	1.4542	0.0877	1.3196	0.0854	1.3615
O-SMC^2	0.0159	0.0134	1.6924	0.0144	0.6852	0.0162	0.4655
	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7
LW	11.8797	16.7328	18.7421	16.8209	16.3860	14.8168	12.3025
Storvik	10.5692	13.3078	15.5469	13.6962	13.8622	11.8743	9.2593
Storvik-CF	5.8181	9.7513	12.5241	9.3262	9.1925	6.7287	5.5044
Storvik-EKF	5.8715	8.4101	9.6863	8.8424	8.8993	7.0859	5.7010
PL	7.0124	9.9405	11.6970	10.0148	10.1936	8.8004	6.7335
O-SMC^2	1.2489	1.5430	1.0991	0.8099	0.9035	0.7052	0.5603

Table 16.16: Mean Monte Carlo Mean Absolute Error (MCMAE) for the parameter and state posterior mean estimation with $n = 50$ runs for different particle filters with the WC98 dataset with $N_p = 5000, N_{obs} = 2304$ and $N_{\boldsymbol{\Phi}} = 2000, N_{\boldsymbol{\theta}} = 100$ for O-SMC^2

16.4.4 Monte Carlo variance

To estimate the mean $MCMAE$, as described in (16.2) and (16.3), LW, Storvik and PL were used with state and parameters priors as specified in Section 16.4.1 on page 231 and the number of particles was $N_p = 5000$. O-SMC^2 was used with $N_{\boldsymbol{\Phi}} = 1000$ and $N_{\boldsymbol{\theta}} = 200$ particles and observation window $h = 250$, using the prior as the importance density (as described in Section 6.2.2 on page 76) for the O-SMC^2 particle filter. The results were then averaged over $n = 50$ runs in order to obtain $\overline{MCMAE}$. These results are illustrated in Figure 16.45 (for LW, Storvik, PL and O-SMC^2), Figure 16.47 (for Storvik-CF/EKF) and summarised in Table 16.16. We also look at the parameter posterior mean and state posterior mean average (and standard errors) at $t = N_{obs}$ for the online filters and compare it to the PMMH state and parameter posterior estimation at $t = N_{obs}$. These results[17] are summarised, respectively, in Tables 16.18 on page 251 and 16.17 on page 250.

[17] Average posterior means can be viewed in Figures B.3 and B.4, in Appendix B.2.

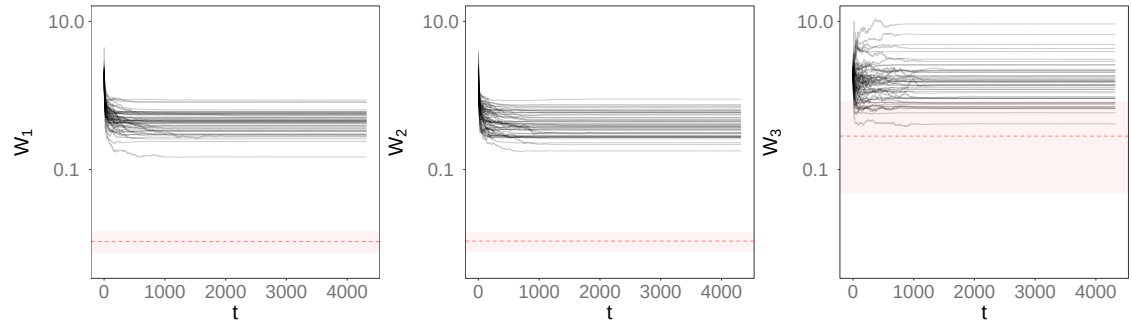
Regarding the variability of the state estimation, the mean $MCMAE$ was also calculated using the PMMH estimate mean as the reference value. As in the previous section, the mean $MCMAE$ will then be the averaged value for n runs.

We can see from Table 16.16 that in terms of parameter estimation variance, O-SMC^2 generally outperforms all the other methods (with the exception of W_3, but still consistent with the other results in this case). From the remaining methods, sufficient-statistics-based ones clearly show less variability than LW. As with the state forecasts, Storvik-CF displays mixed results where the estimation for some parameters (namely $\mathsf{W}_1, \mathsf{W}_2$ and W_6 show less variability than Storvik but not in the remaining parameters (especially in W_3). However, as we can see from Figure 16.45 and 16.47, the estimation for $n = 50$ runs of W_3 shows a high variability when compared with the other parameter suggesting a possible identifiability problem.

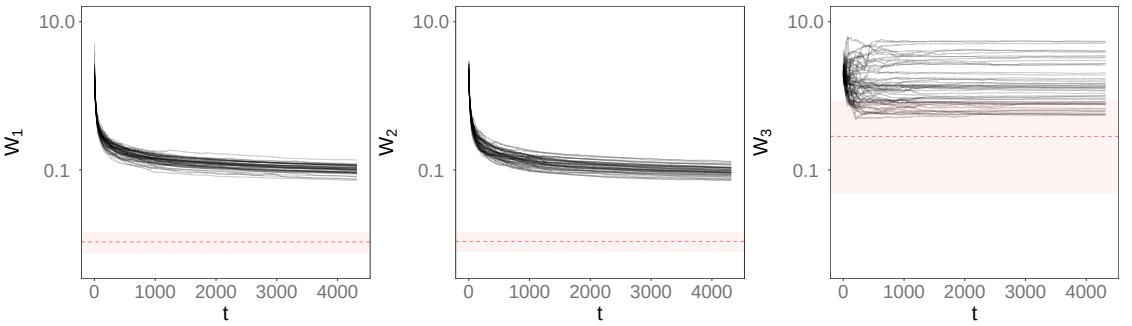
When considering the Monte Carlo variance of the state estimation, we can see similar results. However, in this case, while O-SMC^2 still outperforms the remaining methods, the difference is more marked. Sufficient-statistics-based methods clearly outperform LW. This can be inspected visually from Figures 16.46 and 16.47.

16.4.5 Discrepancies

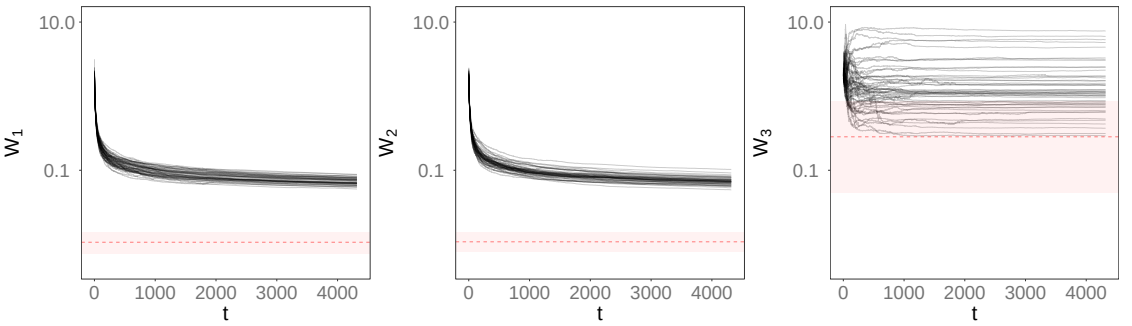
The discrepancy value $d(y_t)$ was used to detect potential anomalies in the data in a online fashion as in previous sections, using the methods described in Section 2.4.3 on page 22. The results are presented in Figures 16.48, 16.49 and 16.50 respectively using LW, Storvik and PL. We can see that none of the filters detected anomalous observations when using a threshold of $d(y_t) > 3$.



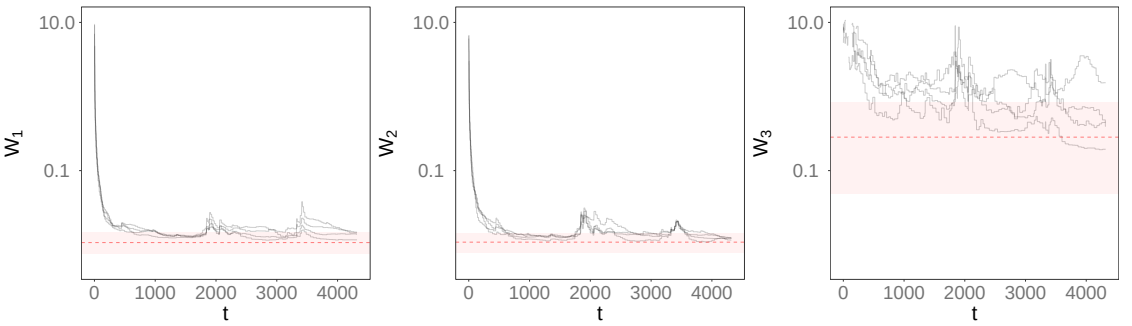
(a) LW estimation history for $\{\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3\}$ using $n = 50$ runs



(b) Storvik estimation history for $\{\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3\}$ using $n = 50$ runs

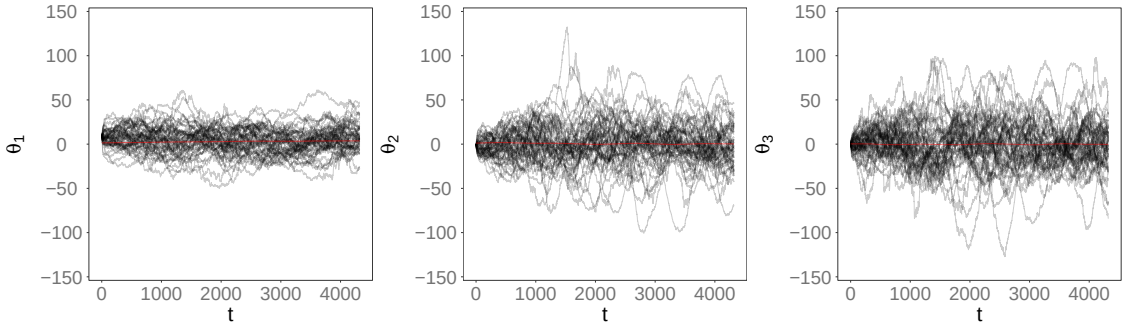


(c) PL estimation for history $\{\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3\}$ using $n = 50$ runs

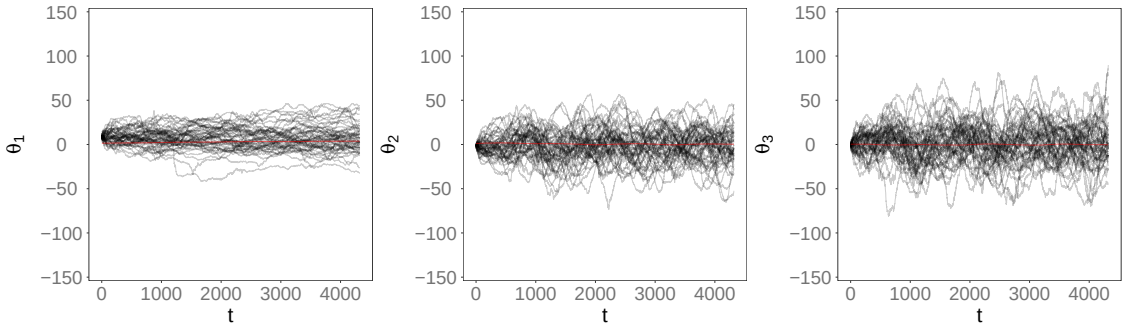


(d) O-SMC^2 estimation history for $\{\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3\}$ using $n = 50$ runs

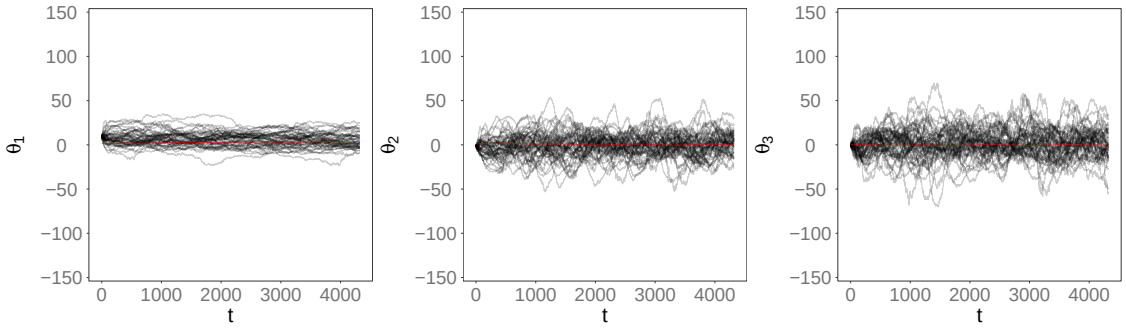
Figure 16.45: Variability in parameter posterior mean estimation history from LW, Storvik, PL and O-SMC^2 for $n = 50$ consecutive runs, using the WC98 data. Horizontal red line represents the PMMH posterior mean and shaded area the 90% equitailed credibility interval(*log-scale*).



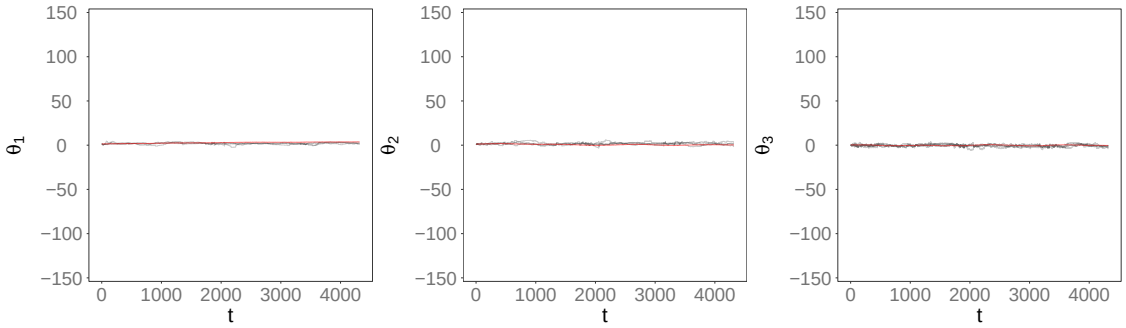
(a) LW estimation history for $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$ using $n = 50$ runs



(b) Storvik estimation history for $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$ using $n = 50$ runs

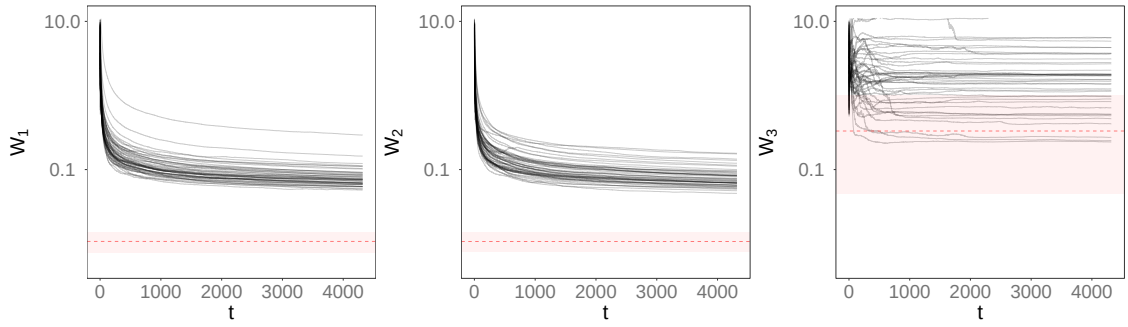


(c) PL estimation for history $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$ using $n = 50$ runs

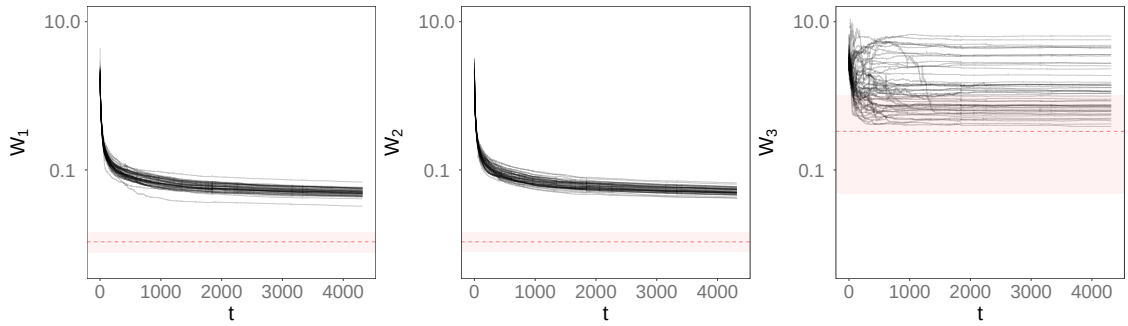


(d) O-SMC^2 estimation history for $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$ using $n = 50$ runs

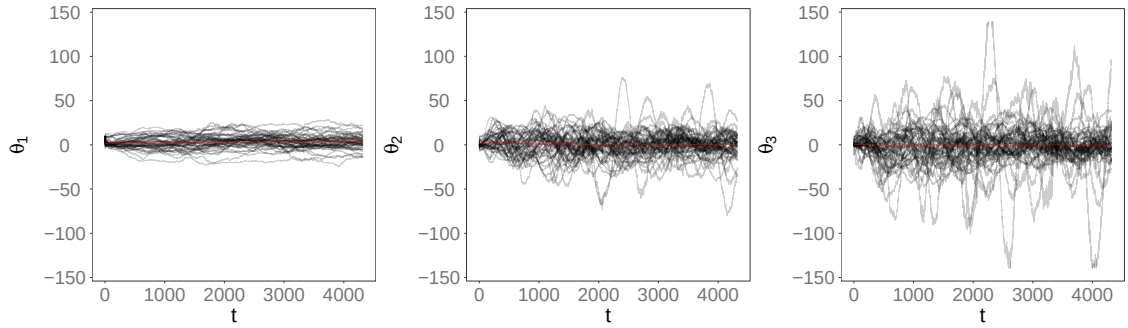
Figure 16.46: Variability in state posterior mean estimation history from LW, Storvik, PL and O-SMC^2 for $n = 50$ consecutive runs, using the WC98 data. Red line represents the PMMH state posterior mean (log-scale).



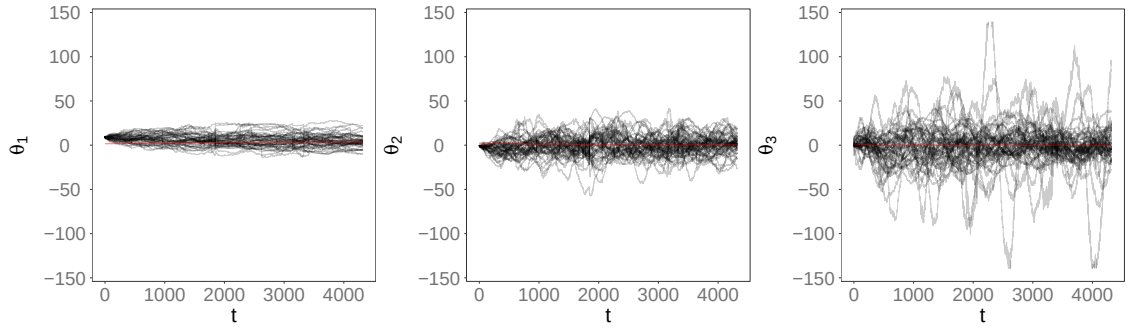
(a) Storvik-CF estimation history for $\{\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3\}$ using $n = 50$ runs



(b) Storvik-EKF estimation history for $\{\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3\}$ using $n = 50$ runs



(c) Storvik-CF estimation history for $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$ using $n = 50$ runs



(d) Storvik-EKF estimation history for $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$ using $n = 50$ runs

Figure 16.47: Variability in parameter (*log-scale*) and state estimation history for Storvik-CF/EKF for $n = 50$ consecutive runs, using the WC98 data

Method	$\overline{\theta}_1$	$\overline{\theta}_2$	$\overline{\theta}_3$	$\overline{\theta}_4$
LW	8.6858 (2.1426)	-3.5011 (3.2814)	-3.5163 (3.803)	1.8948 (3.3442)
Storvik	6.8976 (2.2225)	-1.5058 (2.7843)	3.9171 (3.4773)	-2.2117 (2.6037)
Storvik-CF	2.9414 (1.1658)	0.279 (1.7198)	2.2249 (2.7278)	0.4878 (1.3669)
Storvik-EKF	4.097 (0.9609)	1.0677 (1.2286)	-0.9272 (1.6885)	-0.9148 (1.4858)
PL	4.6599 (1.1723)	2.5428 (1.6095)	-0.4955 (2.3515)	-1.2268 (1.5726)
O-SMC2	2.3014 (0.1118)	1.3366 (0.1383)	-0.7345 (0.1779)	0.4729 (0.1591)
PMMH	3.4554	0.4873	-0.5346	-0.1254
	$\overline{\theta}_5$	$\overline{\theta}_6$	$\overline{\theta}_7$	
LW	7.4492 (2.9243)	-3.0244 (2.6497)	-4.0371 (2.3393)	
Storvik	-4.4401 (2.277)	0.9056 (2.6512)	1.6058 (1.7717)	
Storvik-CF	-1.7742 (1.9245)	0.3778 (1.1203)	0.5643 (1.0407)	
Storvik-EKF	0.2348 (1.4042)	-0.1639 (0.9947)	0.3661 (0.9132)	
PL	2.899 (1.7928)	-1.8775 (1.3975)	-0.8023 (1.1389)	
O-SMC2	-0.1582 (0.1154)	-0.0116 (0.0829)	0.2978 (0.0764)	
PMMH	0.0923	0.0124	0.0847	

Table 16.17: Average state posterior mean (at $t = N_{obs}$) for $n = 50$ runs of the online filters for the WC98 dataset (standard error in brackets) compared to the PMMH state posterior mean estimation.

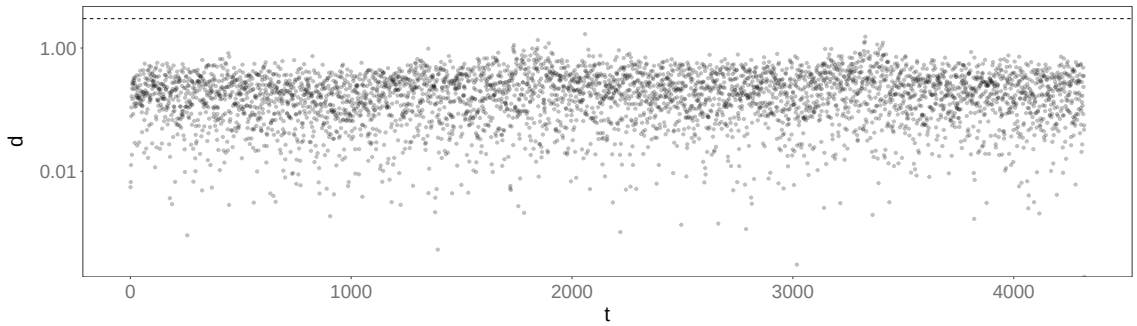


Figure 16.48: Discrepancy values ($d(y_t) > 3$ threshold, log-scale) for the WC98 dataset using the LW filter

Method	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$	$\overline{\mathsf{W}}_4$
LW	0.4553 (0.021)	0.4334 (0.0228)	1.8533 (0.2254)	0.4617 (0.0255)
Storvik	0.1016 (0.0019)	0.0972 (0.002)	1.7707 (0.1893)	0.0987 (0.0017)
Storvik-CF	0.0817 (0.0051)	0.0806 (0.0038)	7.4791 (2.7866)	0.0852 (0.005)
Storvik-EKF	0.0501 (8×10^{-4})	0.0519 (8×10^{-4})	1.7 (0.2152)	0.0511 (8×10^{-4})
PL	0.071 (0.0011)	0.0717 (0.0012)	1.71 (0.2289)	0.072 (0.001)
O-SMC2	0.0132 (2×10^{-4})	0.0133 (2×10^{-4})	0.4807 (0.0393)	0.0135 (8×10^{-4})
PMMH	0.0107	0.0107	0.3266	0.0106
	$\overline{\mathsf{W}}_5$	$\overline{\mathsf{W}}_6$	$\overline{\mathsf{W}}_7$	
LW	2.7205 (0.5604)	0.4429 (0.0179)	2.2766 (0.4433)	
Storvik	1.8039 (0.2815)	0.1002 (0.0019)	1.5703 (0.1476)	
Storvik-CF	2.1287 (0.3142)	0.0785 (0.0035)	1.9128 (0.2629)	
Storvik-EKF	1.6259 (0.2334)	0.0511 (8×10^{-4})	1.4085 (0.1735)	
PL	1.4762 (0.1557)	0.07 (0.0012)	1.4536 (0.1476)	
O-SMC2	0.2687 (0.0167)	0.0133 (8×10^{-4})	0.15 (0.0087)	
PMMH	0.1896	0.0106	0.099	

Table 16.18: Average parameter posterior mean (at $t = N_{obs}$) for $n = 50$ runs of the online filters for the WC98 dataset (standard error in brackets) compared to the PMMH parameter posterior mean estimation.

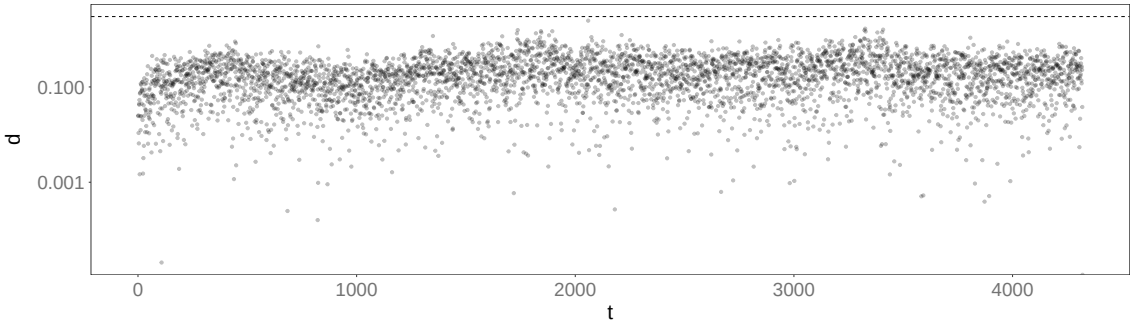


Figure 16.49: Discrepancy values ($d(y_t) > 3$ threshold, log-scale) for the WC98 dataset using the Storvik filter

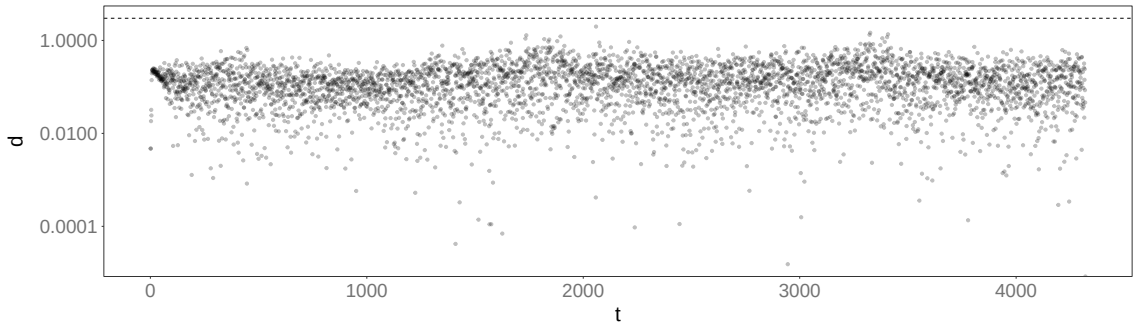


Figure 16.50: Discrepancy values ($d(y_t) > 3$ threshold, log-scale) for the WC98 dataset using the PL filter

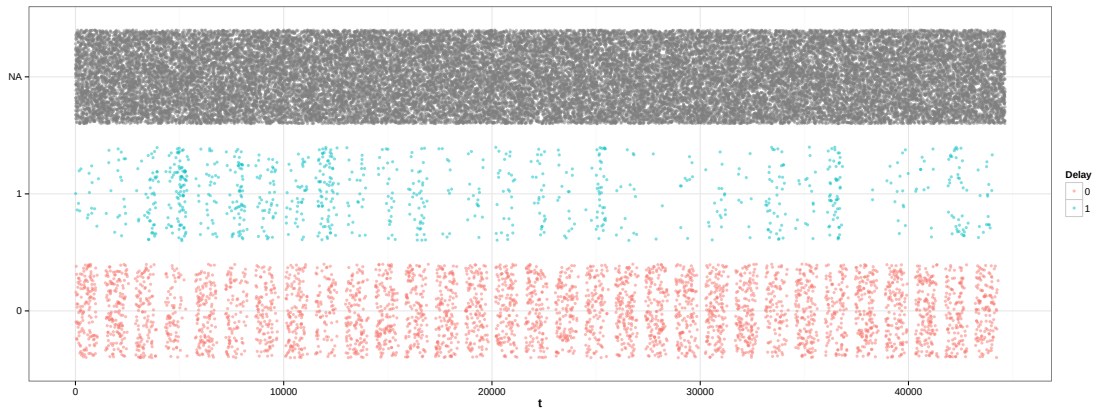


Figure 16.51: Airport delay dataset. On-time flights in red, delayed flights in blue and missing data in grey[18].

16.5 Airport data

The data used for binomial data modelling (the Binomial DLM) comes from the US Department of Transportation's Bureau of Transport Statistics[19] and consist on airport departure times. From the available scheduled and actual departure times we dichotomised the dataset into binary data corresponding to *delayed* and *on-time* flights. A flight was considered delayed if it departed 30 minutes or more after the scheduled time. That is

$$y_t = \begin{cases} 0 & \text{if } |t - t_{expected}| < 30 \\ 1 & \text{if } |t - t_{expected}| \geq 30 \\ \text{NA} & \text{if no flight at } t \end{cases}$$

The data was then converted into a time series with intervals of one minute and since there aren't departures at each time-point t we have inserted missing observations if no departure happened. The airport chosen was the JFK airport in New York City and period was January 2015.

This dataset contains 3822 missing observations, accounting for approximately 88% of the total observations. As we can see from Figure figure 16.51, there is no discernible pattern for the missing data and as such we consider it missing at random.

[18] Jitter applied vertically to all observations to prevent overplotting.

[19] Available at: http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time (Accessed 13^{th} September 2017)

The model chosen was a Binomial DLM as described in (2.25), that is:

$$\begin{aligned} y_t|\eta_t &\sim \text{Binom}\left(\text{logit}^{-1}\{\eta_t\}, n_t\right) \\ \eta_t &= \mathsf{F}^T\boldsymbol{\theta}_t \\ \boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Phi} &\sim \mathcal{N}\left(\mathsf{G}\boldsymbol{\theta}_{t-1}, \mathsf{W}\right). \end{aligned}$$

The dataset corresponds to approximately 3 days worth of minutely data ($N_{obs} = 4320$) for which a locally constant, $\mathcal{P}(1)$ and a Fourier seasonal component with one harmonic[20] and period $p = 1440$ were used, that is

$$\mathcal{M} = \{\mathcal{P}(1), \mathcal{F}(1440, 1)\},$$

which corresponds to the structural matrices

$$\begin{aligned} \mathsf{F} &= [1 \quad 1 \quad 0]^T, \\ \mathsf{G} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\frac{2\pi}{p} & \sin\frac{2\pi}{p} \\ 0 & -\sin\frac{2\pi}{p} & \cos\frac{2\pi}{p} \end{bmatrix}, \\ p &= 1440. \end{aligned}$$

16.5.1 Offline estimation

Offline estimation of the parameter set $\boldsymbol{\Phi} = \{\mathsf{W}\}$ was performed using PMMH, SMC^2 and O-SMC^2. The marginals using PMMH, SMC^2 and O-SMC^2 at time $t = N_{obs}$ are presented in Figure 16.52. The estimation history for SMC^2 and O-SMC^2 is presented respectively in Figure 16.53 on page 257. The traces and auto-correlation plots for PMMH can be viewed in Appendix A.7 on page 289.

SMC^2 and O-SMC^2 were both performed with a number of particles $N_{\boldsymbol{\Phi}} = 2000$ and $N_{\boldsymbol{\theta}} = 1000$ and in the O-SMC^2 case the observation window was $h = 250$ observations[21]. The SIR component of the PMMH used $N_p = 2000$ particles. The priors used with SMC^2,

[20] A number of Storvik runs were performed with different numbers of harmonics to assess the forecast performance. Results are available in Appendix C.3.

[21] The observation window will include $y_{t-250:t}, t > 250$, regardless of observations being "missing" or not.

Method	**Time (s)**	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$
SMC^2	18017.54	0.5763 (0.5346)	0.5264 (0.4421)	0.7022 (0.6075)
O-SMC^2	5627.075	0.5318 (0.5384)	0.4716 (0.4392)	0.6084 (0.7301)
PMMH	–	0.7889 (0.3091)	0.7584 (0.3012)	0.9887 (0.41)

Table 16.19: Summary of computation time and posterior mean estimation (standard deviation in brackets) using offline methods (including O-SMC^2) for the airport dataset.

O-SMC^2 and PMMH for $\boldsymbol{\Phi}_0$ where

$$\mathsf{W}^0 \sim \begin{bmatrix} \mathcal{IG}(1,1) & 0 & 0 \\ 0 & \mathcal{IG}(1,1) & 0 \\ 0 & 0 & \mathcal{IG}(1,1) \end{bmatrix}.$$

The state prior used for SMC^2, O-SMC^2 and PMMH was a normal prior such that

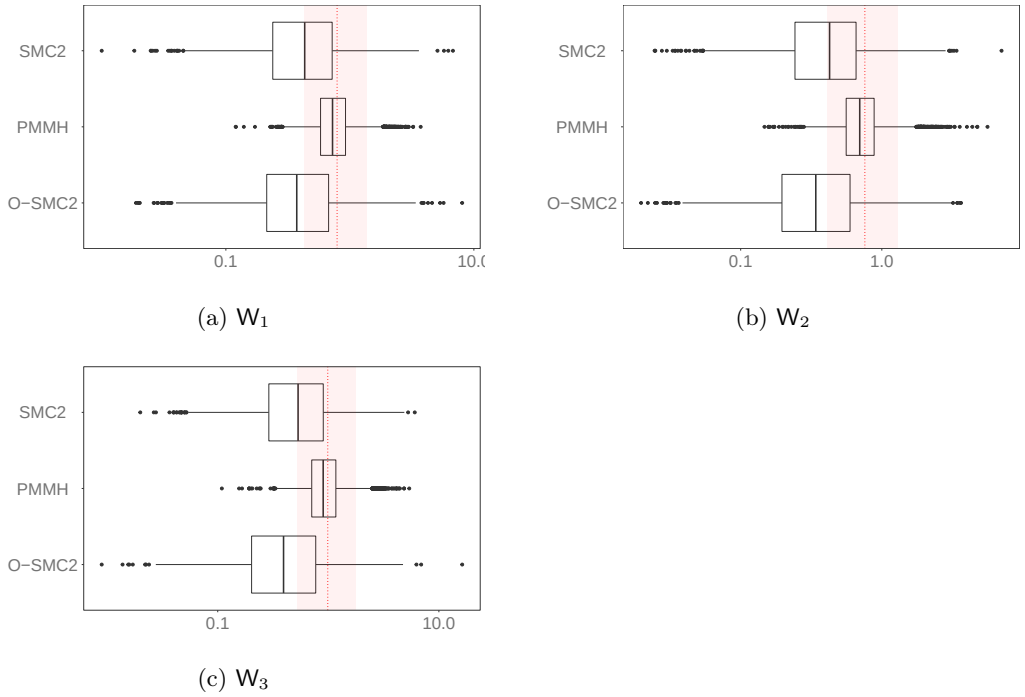
$$\begin{aligned} \boldsymbol{\theta}_0 &\sim \mathcal{N}(\mathbf{m}_0, \mathsf{C}_0) \\ \mathbf{m}_0 &= [0 \quad 0 \quad 0]^T \\ \mathsf{C}_0 &= \operatorname{diag}(40, 40, 40). \end{aligned}$$

A summary of parameter estimation results for the offline methods is presented in Table 16.19. From this table and Figure 16.52 on the following page we can see that the estimated parameters by SMC^2 and O-SMC^2 are consistent, with O-SMC^2 showing a higher accuracy (but only marginally). Both methods show a clear convergence to and overlap with the PMMH estimated posterior as can be seen in Figure 16.52 on the next page.

In terms of computational costs there is a clear difference between the two methods with O-SMC^2 outperforming SMC^2.

16.5.2 Online estimation

State and parameter estimation was performed using the Liu & West, Storvik/EKF and Particle Learning filters, in an online fashion. The Storvik filter was implemented in two variants in this dataset, namely using the prior and the linearised EKF adjusted model as



(a) W_1 (b) W_2

(c) W_3

Figure 16.52: $\boldsymbol{\Phi} = \{\mathsf{W}\}$ parameter posterior estimation using IBIS, O-IBIS (both at $t = N_{obs}$) and PMMH. Vertical red line is the PMMH posterior mean estimation and vertical red bar is the PMMH 90% equitailed credibility interval.

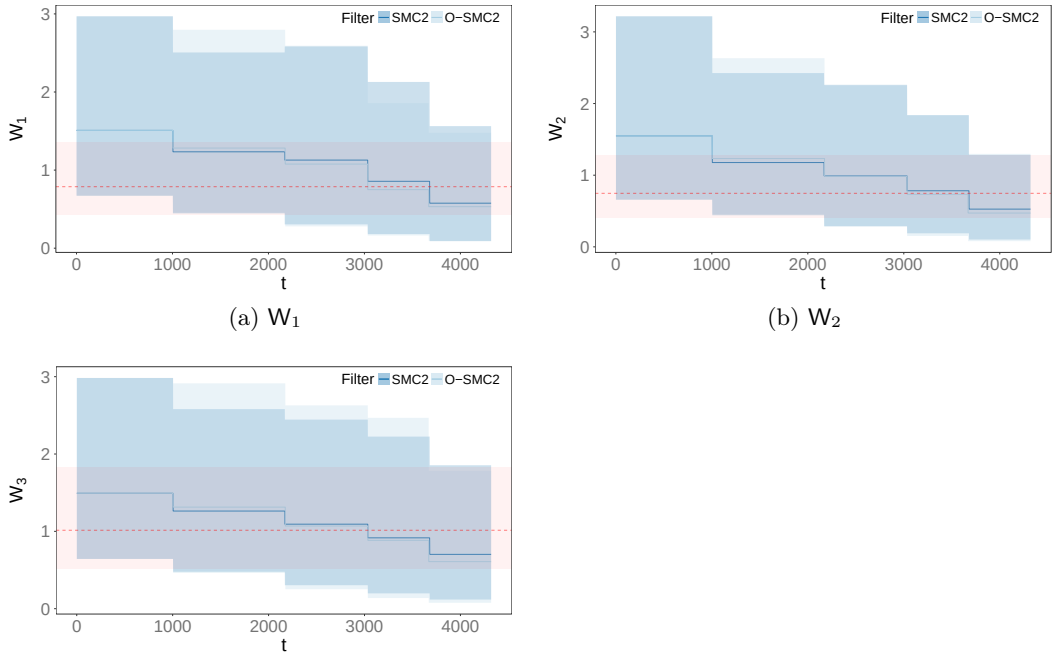
importance densities. The filters were initialised with the same state and parameter priors, drawn respectively from $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{0}, 40\mathbf{I}_3)$. The number of particles used for each filter was $N_p = 4.5 \times 10^4$ and the value chosen after performing several runs[22] with varying N_p. The discount parameter chosen[23] for Liu and West was $\delta = 0.99$. The resampling method used was stratified resampling with a static checkpoint of $n = 1$. The parameters priors were $\mathsf{W}_0 \sim \mathcal{IW}(3, \mathbf{I}_3)$.

In Figure 16.54 we can see the estimated marginals for the parameter set $\boldsymbol{\Phi} = \{\mathsf{W}\}$ using the online methods. The online parameter estimation history results can be seen in Figure 16.55 and a summary of the parameter posteriors at $t = N_{obs}$ can be viewed in Table 16.20.

We can see from Table 16.20 that the sufficient-statistics-based methods outperform LW in terms of parameter estimation when compared to the PMMH result. Specifically, Storvik and PL are consistent with PMMH, whereas Storvik-EKF underestimated the variance components. In terms of computational costs, the values are similar for all methods, which is not surprising considering that due to the fact that 88% of the data is considered

[22]Summary of the full results available in Appendix D.3

[23]The tuning for δ was performed by running several runs of L&W with different values and comparing the state posterior mean MSE and parameter posterior at $t = N_{obs}$. Result available in Appendix E.3.



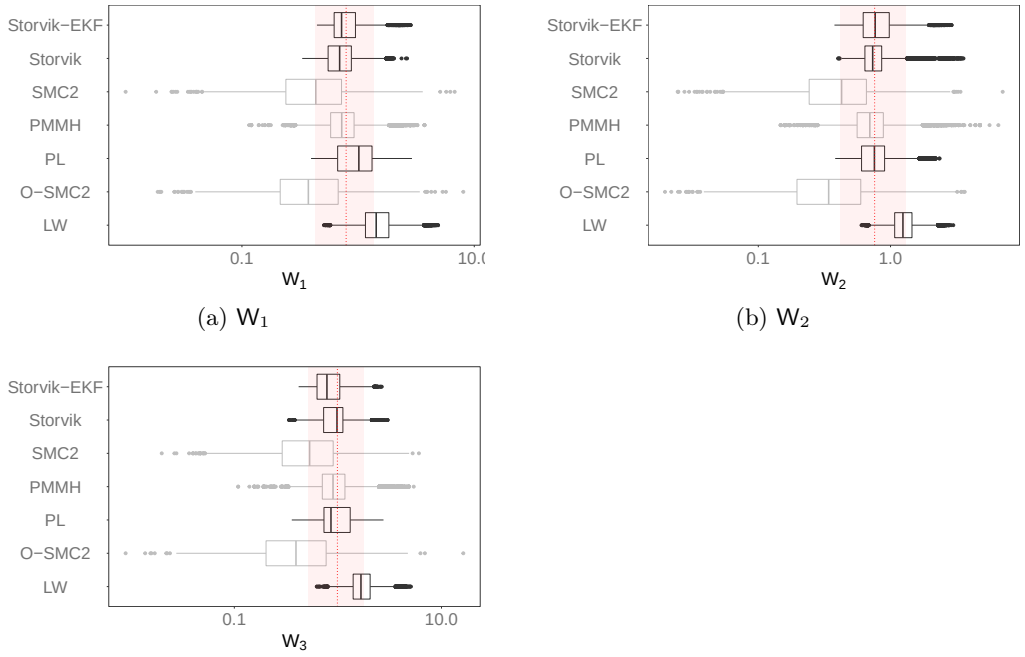
(a) W_1 (b) W_2

(c) W_3

Figure 16.53: $\boldsymbol{\Phi} = \{W\}$ parameter posterior estimation history using SMC^2 and O-SMC^2 for the airport data. Solid lines represent the posterior mean and shaded areas the 90% equitailed credibility interval. PMMH in red (dashed horizontal line is the posterior mean and shaded area the 90% equitailed credibility interval).

"missing", the computation for those steps will be similar (*i.e.* a state update according to the prior with no resampling).

Regarding state estimation, we can see from Table 16.21 that the sufficient-statistics-based methods dominate in terms of state estimation accuracy. In general SMC^2, O-SMC^2 and the remaining sufficient-statistics-based methods show consistent results, with a marginally smaller MSE for SMC^2 and O-SMC^2. LW is the worst performing of the analysed methods for this particular dataset. From the online methods, O-SMC^2 has the highest computational cost. The running time per iteration will vary (but will be computationally bounded) depending on whether we have a rejuvenation step or not, however, the value displayed in Table 16.21 is the mean value for the non-rejuvenating steps at $t > h$, with h as the observation window. We can see this value is several orders of magnitude higher than the remaining methods, and not suitable, for instance, for streaming data with observations arriving every second. However, this computational cost would still be acceptable values for inference in streaming data with a considerably high frequency. We can also see from the state estimation history in Figure 16.56 that posterior variance is marginally smaller for Storvik-EKF.



(a) W_1 (b) W_2

(c) W_3

Figure 16.54: $\boldsymbol{\Phi}$ parameter posterior at $t = N_{obs}$ using the online methods, compared to the offline methods. Vertical red line the posterior mean using PMMH and shaded area the 90% equitailed credibility interval.

16.5.3 Forecast

As in the case of temperature data, a longer k-step ahead forecast ($k = 1440$) was performed on this data. The results for the k-step ahead state forecast are summarised in Table 16.22 and the actual state forecast can be viewed in Figure 16.57.

Regarding state forecast, we can see from Table 16.22 that sufficient-statistics-based methods outperform LW, with Storvik-EKF and PL outperforming Storvik. For the state vector component θ_3 (the highest frequency harmonic of the seasonal component), Storvik-EKF is the most accurate when compared to PMMH.

16.5.4 Monte Carlo variance

Similarly to the previous sections, to estimate the mean $MCMAE$, as described in (16.2) and (16.3), LW, Storvik and PL were used with state priors as specified in Section 16.5,

[24] Average time of non-resampling steps with $t > k$.
[25] Average time of non-resampling steps with $t > k$.

Method	**Time (s)**	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$
LW	244.508	1.5496 (0.5535)	1.2956 (0.3262)	1.7875 (0.5454)
Storvik	292.798	0.7563 (0.2708)	0.7942 (0.2683)	0.952 (0.2828)
Storvik-EKF	370.296	0.8102 (0.267)	0.8194 (0.2539)	0.8892 (0.3494)
PL	346.982	1.0492 (0.4221)	0.7999 (0.2673)	1.0243 (0.4307)
PMMH	–	0.7804 (0.3164)	0.7331 (0.3132)	0.9789 (0.4231)

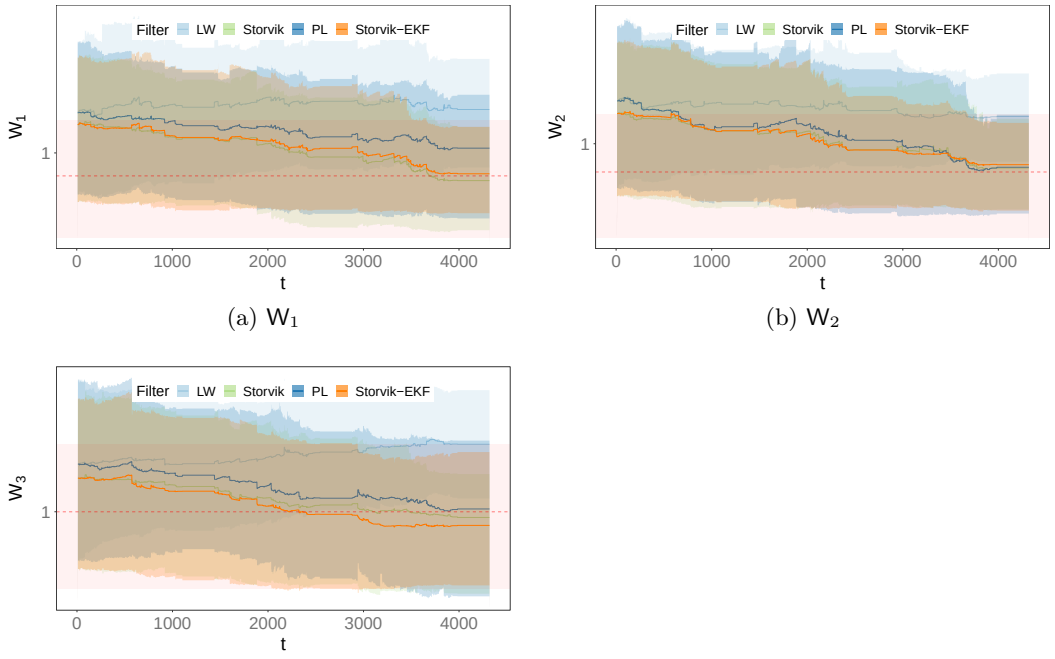
Table 16.20: Computation time, parameter posterior mean estimation (and standard deviation, in brackets) at $t = N_{obs}$ with online methods for the airport data.

Method	**MSE**			**Time**	
	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	*iteration (ms)*	*total (s)*
LW	59.2434	42.1278	48.9085	56.59	244.508
Storvik	30.1322	30.6581	40.0322	67.7773	292.798
Storvik-EKF	32.6937	28.7214	37.8335	85.7167	370.296
PL	33.1388	33.284	36.1698	80.3199	346.982
SMC^2	26.406	27.6547	35.4881	3078.0[24]	18017.54
O-SMC^2	22.8451	22.2972	22.1871	2314.5[25]	5627.075

Table 16.21: State posterior mean MSE (relatively to PMMH state posterior mean estimation) and computation time with different particle filters for the airport dataset.

Method	**MSE**		
	$\hat{\boldsymbol{\theta}}_{1,t+k}$	$\hat{\boldsymbol{\theta}}_{2,t+k}$	$\hat{\boldsymbol{\theta}}_{3,t+k}$
LW	331.59	168.4	121.98
Storvik	196.12	143.7	102.24
Storvik-EKF	188.46	110.4	74.437
PL	170.12	123.92	85.142

Table 16.22: State posterior mean k-step ($k = 1440$) ahead forecast MSE (relatively to PMMH state posterior mean estimation) for different particle filters with the airport dataset.



(a) W_1 (b) W_2

(c) W_3

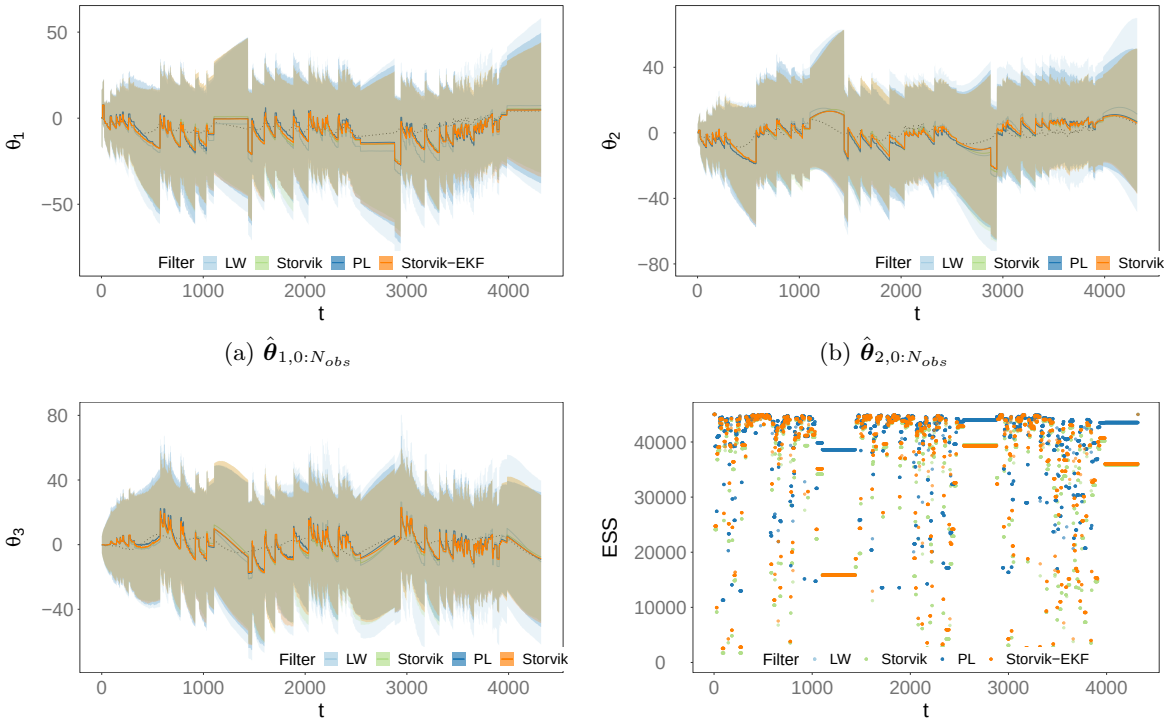
Figure 16.55: Parameter posterior estimation history for $\boldsymbol{\Phi}$ using online methods for the airport data. Solid lines represents the posterior mean, shaded red area the 90% equitailed credibility intervals. Dashed horizontal line represents PMMH posterior mean and shaded area the 90% equitailed credibility interval.

parameter priors

$$\mathsf{W}_0 \sim \begin{bmatrix} \mathcal{IG}(5,6) & 0 & 0 \\ 0 & \mathcal{IG}(5,6) & 0 \\ 0 & 0 & \mathcal{IG}(5,6) \end{bmatrix}$$

and the number of particles was $N_p = 5000$. O-SMC^2 was used with $N_{\boldsymbol{\Phi}} = 2000$ and $N_{\boldsymbol{\theta}} = 250$ particles and observation window $h = 1000$, using the prior as the importance density (as described in Section 6.2.2 on page 76) for the log-likelihood estimator SIR. The observation window for O-SMC^2 was larger than in previous sections to accommodate for high number of missing observation. As previously, the results were then averaged over $n = 50$ runs in order to obtain $\overline{MCMAE}$. The results for state estimation variability are illustrated in Figure 16.59 and for the parameter variability in Figure 16.58. We also look at the parameter posterior mean and state posterior mean average (and standard errors) at $t = N_{obs}$ for the online filters and compare it to the PMMH state and parameter posterior estimation at $t = N_{obs}$. These results[26] are summarised, respectively, in Tables 16.25 and

[26] Average posterior means can be viewed in Figures B.5 and B.6, in Appendix B.3.



(a) $\hat{\boldsymbol{\theta}}_{1,0:N_{obs}}$

(b) $\hat{\boldsymbol{\theta}}_{2,0:N_{obs}}$

(c) $\hat{\boldsymbol{\theta}}_{3,0:N_{obs}}$

(d) Effective Sample Size

Figure 16.56: State posterior estimation using different particle filters on the airport dataset and Effective Sample Size. Solid colour lines represent state posterior mean, shaded area 90% equitailed credibility intervaland dashed black line PMMH state posterior mean.

16.24.

Regarding the variability of the state estimation, the mean $MCMAE$ was also calculated using the PMMH estimate mean as the reference value. As in the previous section, the mean $MCMAE$ will then be the averaged value for n runs.

We can see from Table 16.16 that in terms of parameter estimation variance, O-SMC^2 generally outperforms all the other methods (with the exception of W_3, but still consistent with the other results in this case). From the remaining methods, sufficient-statistics-based ones clearly show less variability than LW. Within the sufficient-statistics-based methods, Storvik-EKF displays the lowest variability, even outperforming PL.

When considering the Monte Carlo variance of the state estimation, we can see similar results. However, in this case, while O-SMC^2 still outperforms the remaining methods, the difference is more marked. Sufficient-statistics-based methods clearly outperform LW. This can be inspected visually from Figures 16.46 and 16.47.

Method	$\overline{MCMAE}$					
	W_1	W_2	W_3	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$
LW	0.9998	1.0035	0.9923	8.0133	7.6007	8.3083
Storvik	0.3147	0.3505	0.2511	4.8178	4.7752	5.7318
Storvik-EKF	0.3209	0.3142	0.2083	5.0508	4.7619	5.9026
PL	0.4618	0.4478	0.3587	6.2117	5.8276	6.8858
O-SMC^2	0.3955	0.2643	0.3727	3.2762	3.5579	3.5228

Table 16.23: Mean Monte Carlo Mean Absolute Error (MCMAE), between particle filters and PMMH state and parameter posterior mean with $n = 50$ runs for the airport dataset with $N_p = 5000, N_{obs} = 2304$ and $N_{\boldsymbol{\Phi}} = 2000, N_{\boldsymbol{\theta}} = 250$ for O-SMC^2.

Method	$\overline{\boldsymbol{\theta}}_1$	$\overline{\boldsymbol{\theta}}_2$	$\overline{\boldsymbol{\theta}}_3$
LW	7.4904 (1.1782)	8.2949 (1.2437)	-12.3185 (1.2494)
Storvik	4.5314 (0.4042)	6.162 (0.4817)	-9.7284 (0.3806)
Storvik-EKF	5.914 (0.5682)	6.5569 (0.5264)	-8.1812 (0.5302)
PL	5.4962 (0.5157)	5.2959 (0.5287)	-9.4284 (0.4995)
O-SMC2	-8.7347 (0.3347)	-13.0508 (0.4918)	-11.5654 (0.4603)
PMMH	4.0508	4.4866	-9.38

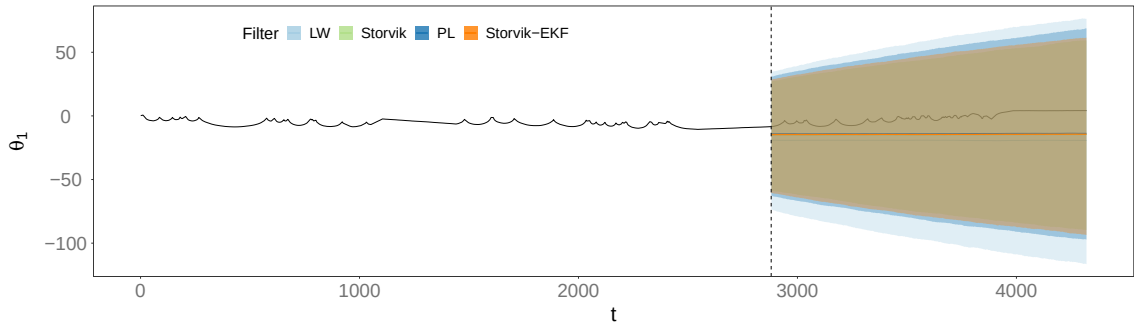
Table 16.24: Average state posterior mean (at $t = N_{obs}$) for $n = 50$ runs of the online filters for the airport dataset (standard error in brackets) compared to the PMMH state posterior mean estimation.

Method	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$
LW	1.8936 (0.0963)	1.8634 (0.0866)	2.191 (0.1153)
Storvik	0.923 (0.0313)	0.9243 (0.0323)	1.0998 (0.0497)
Storvik-EKF	0.9571 (0.0333)	0.8486 (0.0286)	1.0201 (0.0349)
PL	1.0506 (0.0366)	1.0238 (0.0411)	1.1899 (0.0366)
O-SMC2	0.8945 (0.0218)	0.845 (0.0233)	1.0098 (0.0274)
PMMH	0.7932	0.7647	0.9995

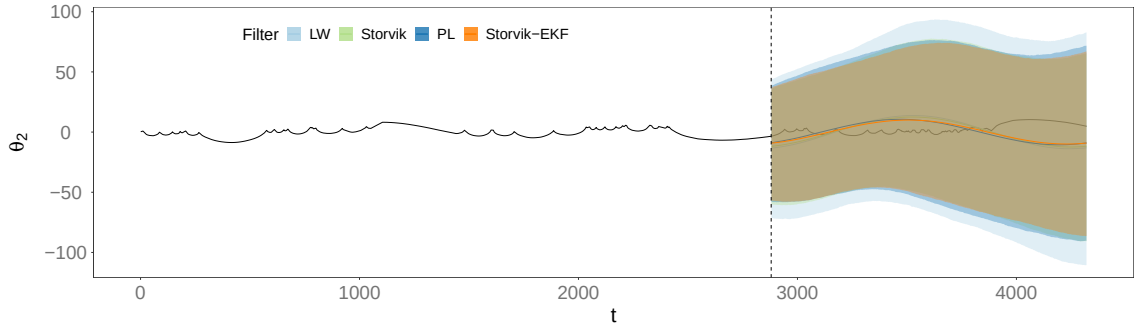
Table 16.25: Average parameter posterior mean (at $t = N_{obs}$) for $n = 50$ runs of the online filters for the airport dataset (standard error in brackets) compared to the PMMH parameter posterior mean estimation.

16.5.5 Discrepancies

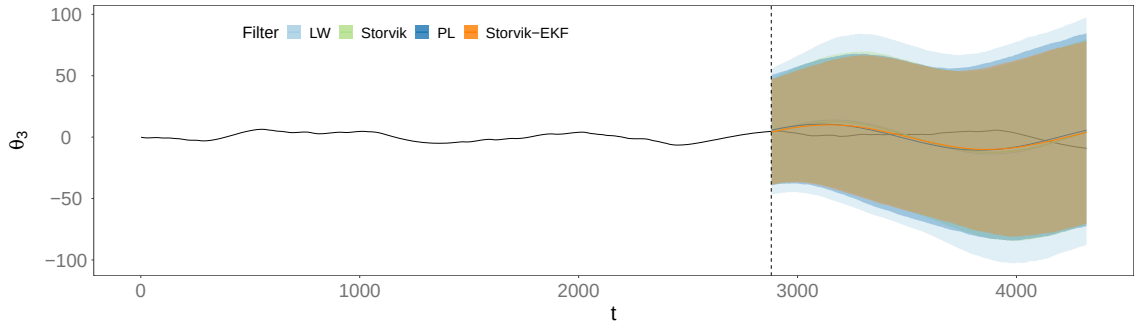
As in previous sections, the discrepancy value $d(y_t)$ was used to detect potential anomalies in the data in a online fashion as in previous sections, using the methods described in Section 2.4.3. The results are presented in Figures 16.61, 16.62 and 16.63 respectively using LW, Storvik and PL. The discrepancy value was only calculated for time-points where the observation was not missing and we can see that none of the filters detected anomalous observations when using a threshold of $d(y_t) > 3$.



(a) θ_1

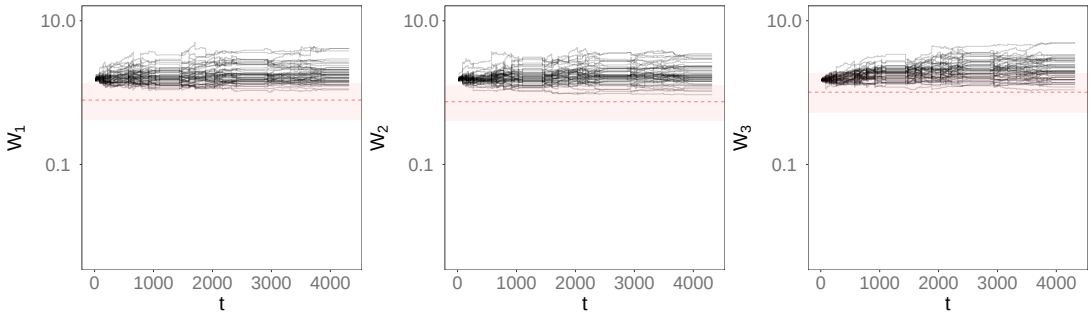


(b) θ_2

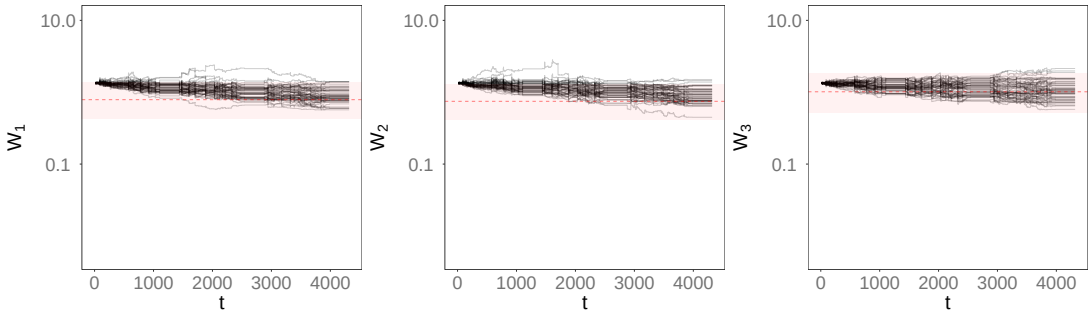


(c) θ_3

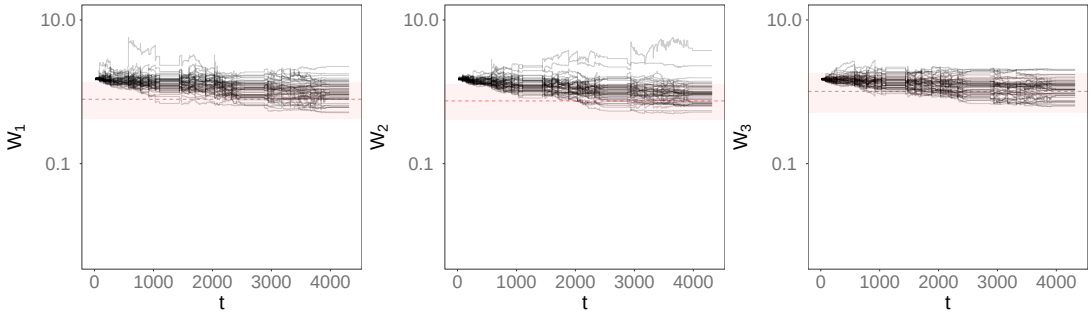
Figure 16.57: State component k-step ahead forecast on the temperature dataset ($k = 1440$). Solid colour line represent the state forecast posterior mean, shaded areas the 90% equitailed credibility interval. Solid black line represent the PMMH state posterior mean estimation.



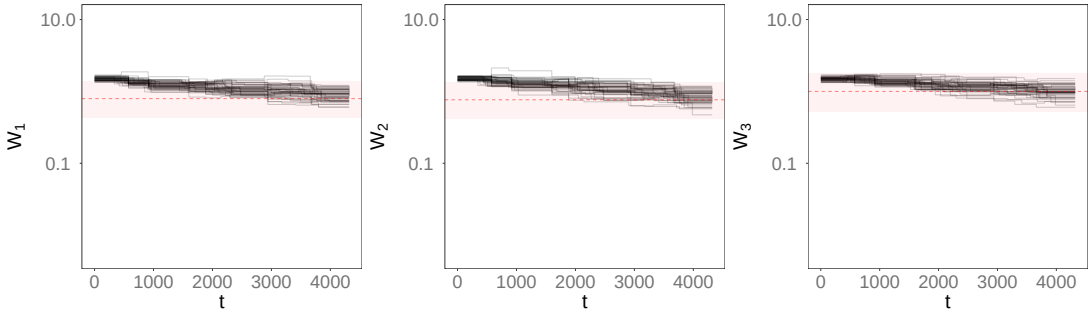
(a) LW parameter posterior mean estimation history for $\{\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3\}$ using $n = 50$ runs.



(b) Storvik parameter posterior mean estimation history for $\{\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3\}$ using $n = 50$ runs.

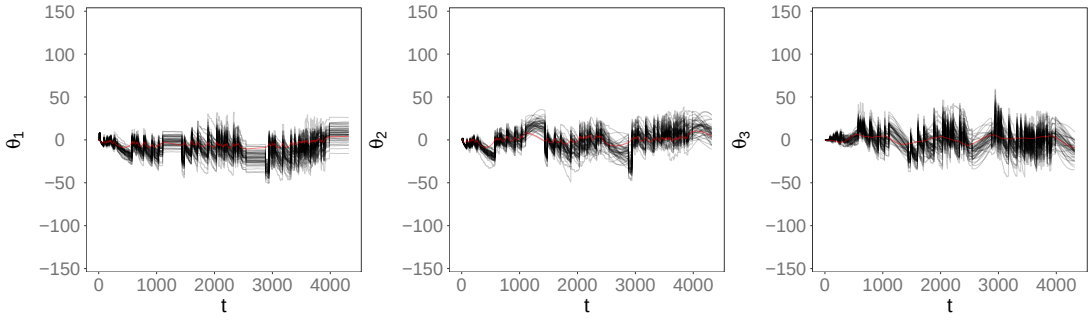


(c) PL parameter posterior mean estimation history for $\{\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3\}$ using $n = 50$ runs.

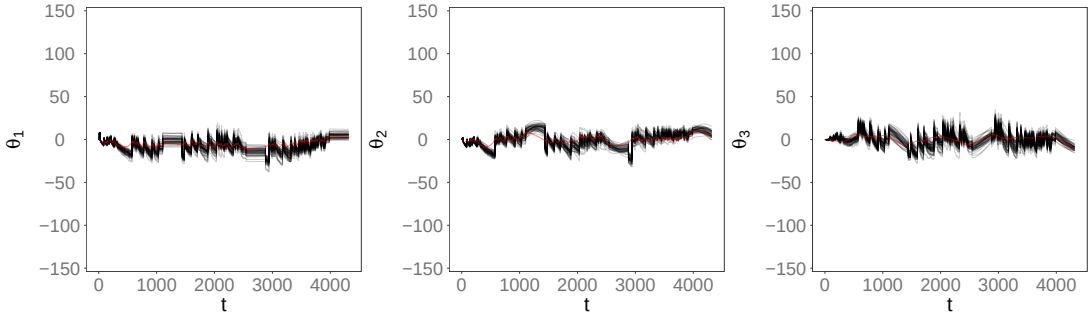


(d) O-SMC^2 parameter posterior mean estimation history for $\{\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3\}$ using $n = 50$ runs.

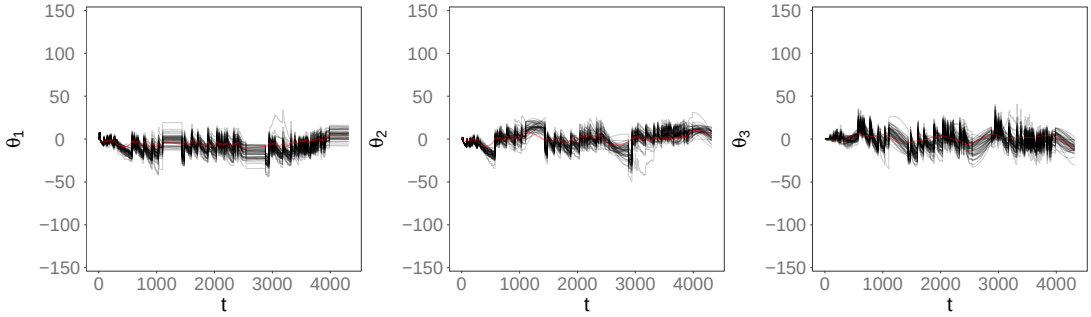
Figure 16.58: Variability in parameter estimation history from LW, Storvik, PL and O-SMC^2 for $n = 50$ consecutive runs, using the airport data. Red horizontal line and shaded area represent, respectively, PMMH parameter posterior mean and 90% equitailed credibility interval(*log-scale*).



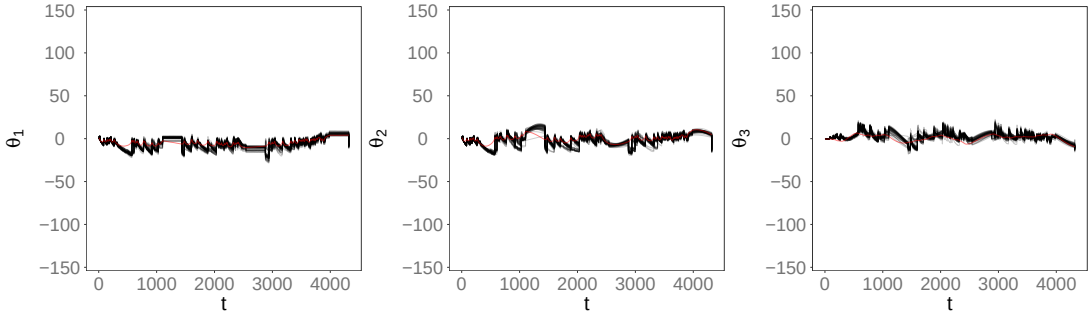
(a) LW state posterior mean estimation history for $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$ using $n = 50$ runs.



(b) Storvik state posterior mean estimation history for $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$ using $n = 50$ runs.

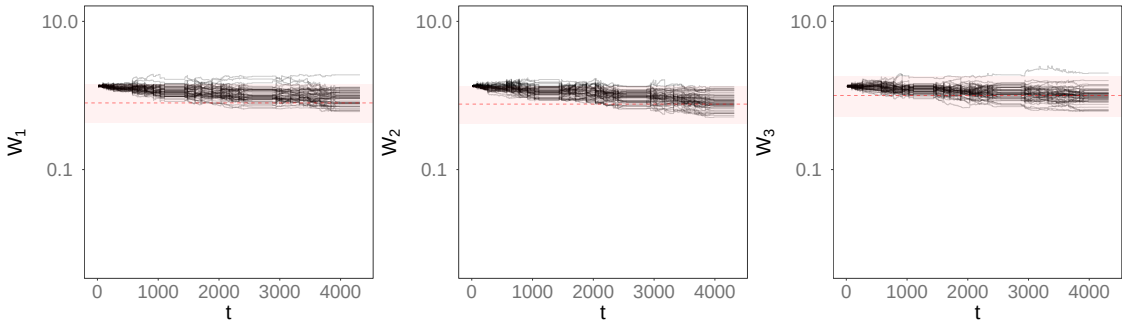


(c) PL state posterior mean estimation for history for $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$ using $n = 50$ runs.

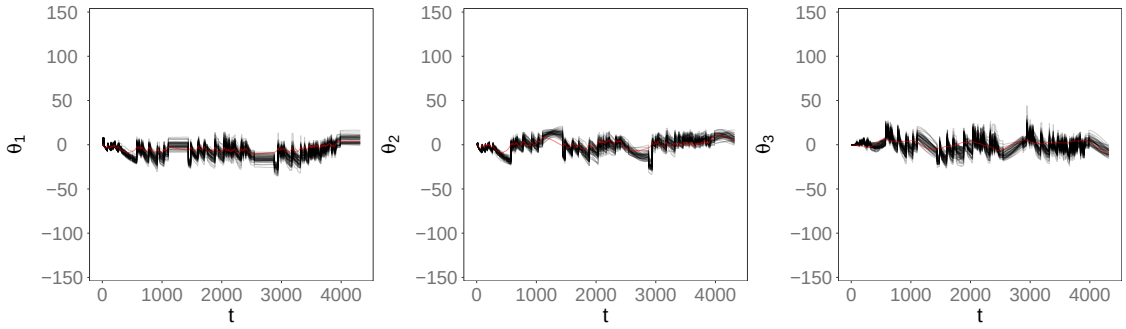


(d) O-SMC^2 state posterior mean estimation history for $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$ using $n = 50$ runs.

Figure 16.59: Variability in state posterior mean estimation history with LW, Storvik, PL and O-SMC^2 for $n = 50$ consecutive runs, using the airport data (*log-scale*). Red line represents PMMH state posterior mean.



(a) Storvik-EKF parameter posterior mean estimation history for $\{\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3\}$ using $n = 50$ runs



(b) Storvik-EKF state posterior mean estimation history for $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$ using $n = 50$ runs

Figure 16.60: Variability in state and parameter posterior mean estimation history using Storvik-EKF for $n = 50$ consecutive runs, for the airport data (*log-scale*). Red vertical line represents PMMH parameter posterior mean and shaded area 90% equitailed credibility interval(*top*). Red line represents PMMH state posterior mean (*bottom*).

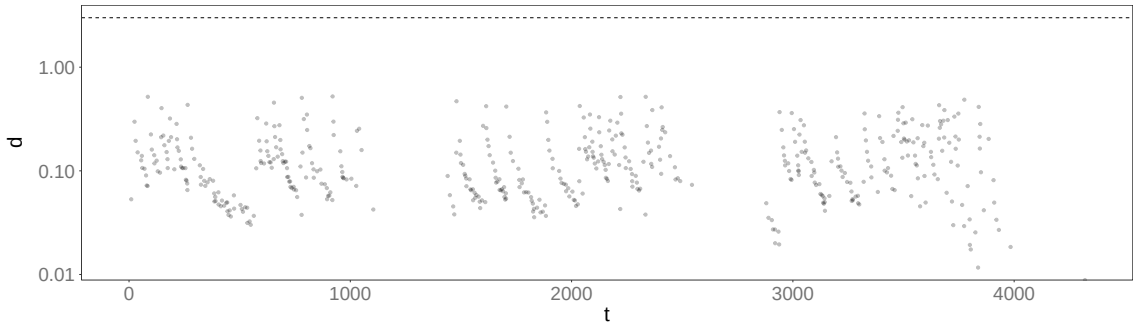


Figure 16.61: Discrepancy values ($d(y_t) > 3$ threshold, log-scale) for the airport dataset using the LW filter

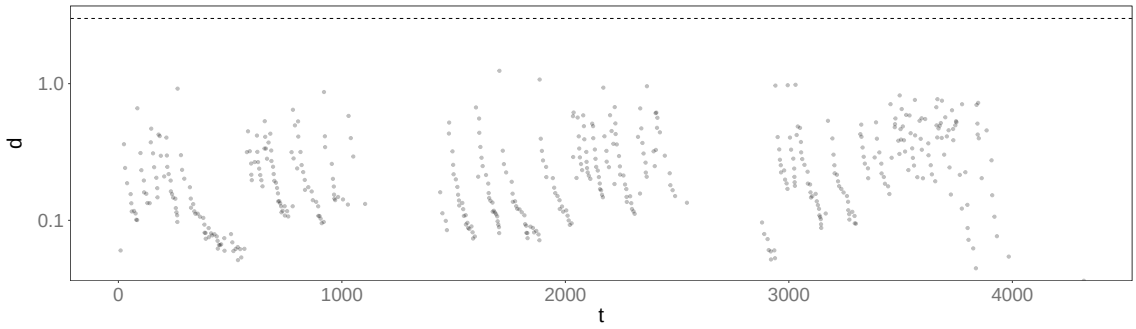


Figure 16.62: Discrepancy values ($d(y_t) > 3$ threshold, log-scale) for the airport dataset using the Storvik filter

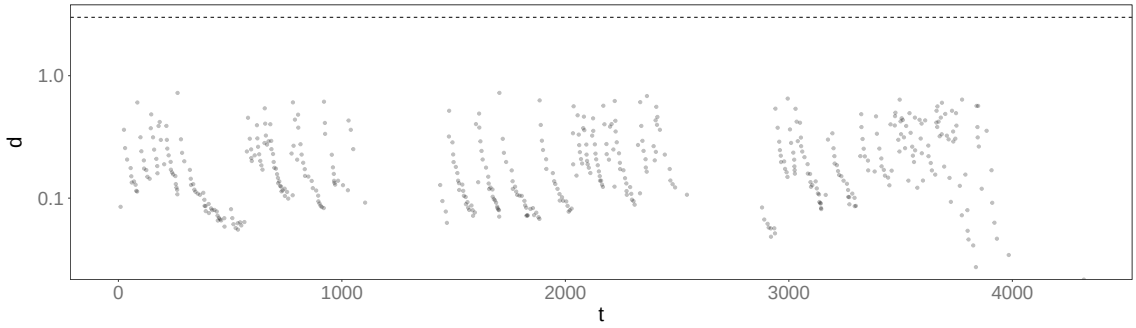


Figure 16.63: Discrepancy values ($d(y_t) > 3$ threshold, log-scale) for the airport dataset using the PL filter

Chapter 17

Conclusions

In this thesis we analysed several methods available for online state and parameter estimation in DGLMs with the aim of performing statistical inference in near real-time streaming data scenarios.

The main novel contribution is the application of two classes of proposals for non-linear DGLM importance sampling. One of the proposals, based on a Linear Bayes linearisation of the model (Section 6.2.4 on page 79), has not, to our knowledge, been applied to SMC in the context of state or both state and parameter estimation. The other proposal, based on a linearised adjusted model (Section 6.2.5 on page 80), although applied in SMC in the context of state estimation, has not been previously applied to both state and parameter estimation, especially in sufficient-statistics-based algorithms. We have also contributed with an *ad-hoc* formulation for IBIS in DLMs (Section 14.1 on page 171) and SMC^2 in non-linear DGLMs (Section 15.1 on page 184), which allow for online state and parameter estimation for a approximated target distribution. The IBIS/O-IBIS methods were also implemented using an SVD implementation for the KF recursions (Section 4.1 on page 36) in order to investigate potential improvements in numerical accuracy which might result from using real-world datasets.

Furthermore, we have analysed all these methods with real-world data which tried to typify possible streaming data sources, that is, continuous and discrete data with structures including seasonal patterns, presence of outliers and missing data with the purpose of comparing the different methods in terms of state and parameter estimation, forecast performance and computational cost. In the examples presented in this thesis, all with simple models, simulated data and a high number of particles, we have seen that the online methods present similar results in terms of state and parameter estimation. However, when using real data and under computational budget constraints, this is not the case. From our results we observed that, in general, within sequential methods, SMC^2 and O-SMC^2

outperform all the remaining methods in terms of state and parameter estimation accuracy.

For online methods, O-SMC^2 provided the best estimates when compared with our "gold standard", PMMH.

For the remaining methods, sufficient-statistics-based methods outperformed state augmentation approaches. Within sufficient-statistics methods, Particle Learning generally provided better results than Storvik, even when not using the "full" formulation as in the non-linear DGLMs, *i.e.* marginalising the state by using an essential state vector.

Some of the topics relevant to near real-time estimation of streaming data were also analysed, such as the behaviour of these algorithms for long-running time-series and how Monte Carlo errors introduced by the discrete approximation of the posteriors would affect the estimation. We have seen that, although sufficient-statistics clearly reduce the variability of the estimates when compared to state augmentation, the rejuvenation steps of SMC^2/O-SMC^2 provide the best results regarding this criteria.

The increased accuracy (when compared to PMMH) of O-SMC^2 comes with an increased computational cost. While sufficient-statistics methods can be used for higher-frequency data (with computational costs around a few hundred milliseconds per iteration for a considerable amount, $N_p = 4.5 \times 10^4$, of particles), O-SMC^2 increases that cost to the order of a few seconds per iteration in the models analysed. However, this increase would still be perfectly suitable for some of the data streams analysed, where the highest frequency arrival of data was of one observation per minute.

Although SMC^2 could be presented as a competitor to traditional MCMC methods in some scenarios, this is not the case for the other SMC-based methods analysed.

However, for the purposes of inference such as short to medium range state and observation forecast as well as anomaly detection, these methods provided a useful set of tools for when accuracy is not paramount, but a trade-off between performance and accuracy is desirable. The online methods presented provide, for instance, a straight-forward framework in which the number of particles could be adjusted between iterations to reduce the computational burden in a real-time fashion at the cost of a lower accuracy and higher variability.

We have also analysed the behaviour of these methods with anomalous observations, where we could observe that rejuvenation-based and state-augmentation methods could not avoid collapse, while sufficient-statistics-based methods avoided particle collapse but provided poor forecast results. Moreover, as we could see, methods such as the discrepancy calculation provide a way to perform anomaly detection in an online way. As such, in real world applications, where arguably targeting an approximated posterior is preferable to a total filter collapse, if, at any iteration, an anomaly is detected, this step could simply

be repeated with the observation removed and simply updating the latent states without applying a correction step.

The online methods displayed ability to cope with a dataset where the majority of the observations were missing, showing consistency with the PMMH estimation.

Using the alternative proposals presented in this thesis, we could generally see an increase in state estimation accuracy and a lower variability for long-running estimations, especially when using the linearised adjusted model.

Future work could include the inclusion of different models to deal with potential anomalous observations, such as replacing a Poisson DLM with a Negative Binomial DLM and even the possible inclusion of a non-DGLM, such as a t distribution based model as a replacement for the DLM.

A deeper investigation of the behaviours of the O-SMC^2/O-IBIS with regards to dynamic particle numbers and different kernels (such as Particle Gibbs) could be warranted.

A formal theoretical framework to determine convergence properties, if any, of these methods is still required, although beyond the scope of the work in this thesis.

Bibliography

Andrieu, C., Doucet, A. & Holenstein, R. 2009 Particle Markov Chain Monte Carlo for Efficient Numerical Simulation. *Monte Carlo and Quasi-Monte Carlo Methods 2008* **1** (Mcmc), 369–382.

Andrieu, C., Doucet, A. & Holenstein, R. 2010 Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **72** (3), 269–342.

Andrieu, C. & Roberts, G. O. 2009 The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics* **37** (2), 697–725.

Arlitt, M. & Jin, T. 2000 A workload characterization study of the 1998 World Cup Web site. *IEEE Network* **14** (3), 30–37.

Arulampalam, M. S., Maskell, S., Gordon, N. & Clapp, T. 2002 A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing* **50** (2), 174–188.

Brooks, S., Gelman, A., Jones, G. & Meng, X.-L. 2011 *Handbook of Markov Chain Monte Carlo*. CRC Press/Taylor & Francis.

Cappé, O., Moulines, E. & Ryden, T. 2006 *Inference in Hidden Markov Models*, , vol. 48.

Carpenter, J., Clifford, P. & Fearnhead, P. 1999 Improved particle filter for nonlinear problems. *Radar, Sonar and Navigation, IEE Proceedings* - **146** (1), 2–7.

Carvalho, C. M., Johannes, M. S., Lopes, H. F. & Polson, N. G. 2010 Particle Learning and Smoothing. *Statistical Science* **25** (1), 88–106.

Chopin, N. 2002 A sequential particle filter method for static models. *Biometrika* **89** (3), 539–551.

Chopin, N. 2004 Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Annals of Statistics* **32** (6), 2385–2411.

Chopin, N., Iacobucci, A., Marin, J.-M., Mengersen, K., Robert, C. P., Ryder, R. & Schäfer, C. 2010 On Particle Learning. *Review Literature And Arts Of The Americas* **1** (2008), 14.

Chopin, N., Jacob, P. E. & Papaspiliopoulos, O. 2013 SMC2: An efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **75** (3), 397–426.

Chopin, N., Ridgway, J., Gerber, M. & Papaspiliopoulos, O. 2015 Towards automatic calibration of the number of state particles within the SMC$ˆ2$ algorithm **2** (1), 1–18.

Creal, D. 2012 A Survey of Sequential Monte Carlo Methods for Economics and Finance. *Econometric Reviews* **31** (3), 245–296.

Crisan, D. & Doucet, A. 2002 A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing* **50** (3), 736–746.

Del Moral, P., Doucet, A. & Jasra, A. 2006 Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **68** (3), 411–436.

Dobson, A. 2000 *An Introduction to Generalized Linear Models*. CRC press.

Douc, R., Cappe, O. & Moulines, E. 2005 Comparison of Resampling Schemes for Particle Filtering. *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis* pp. 64–69.

Doucet, A., de Freitas, N. & Gordon, N. E. 2001*a* Sequential Monte Carlo Methods in Practice. *Technometrics* p. 583.

Doucet, A., Godsill, S. & Andrieu, C. 2000 On sequential Monte Carlo sampling methods for Bayesian filtering.

Doucet, A., Gordon, N. J., Krishnamurthy, V. & Member, S. 2001*b* Particle Filters for State Estimation of Jump Markov Linear Systems. *Ieee Transactions on Signal Processing* **49** (3), 613–624.

Fahrmeir, L. 1992 Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association* **87** (418), 501–509.

Fearnhead, P. 2002 Markov chain Monte Carlo, sufficient statistics, and particle filters. *Journal of Computational and Graphical Statistics* .

Fearnhead, P. & Taylor, B. M. 2013 An adaptive sequential monte carlo sampler. *Bayesian Analysis* **8** (2), 411–438.

Gamerman, D. 1998 Markov chain Monte Carlo for dynamic generalised linear models. *Biometrika* **85** (1), 215–227.

Gilks, W. R. & Berzuini, C. 2001 Following a moving target-Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63** (1), 127–146.

Gordon, N., Salmond, D. & a.F.M. Smith 1993 Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F Radar and Signal Processing* **140** (2), 107.

Hans, R. K. 2003 Recursive Monte Carlo Filters : Algorithms and Theoretical Analysis. *Annals of Statistics* .

Hol, J. D., Schön, T. B. & Gustafsson, F. 2006 On resampling algorithms for particle filters. *NSSPW - Nonlinear Statistical Signal Processing Workshop 2006* .

Jazwinski, A. 1966 Filtering for nonlinear dynamical systems. *IEEE Transactions on Automatic Control* **11** (4), 765–766.

Jazwinski, A. H. 1970 *Stochastic Processes and Filtering Theories*, , vol. 1, Academi. Dover Publications.

Julier, Â. & Uhlmann, Â. 1997 New extension of the Kalman filter to nonlinear systems. *Proc. SPIE Vol. 3068* **3068**, 182–193.

Kalman, R. E. 1960 A new approach to linear filtering and prediction problems. *Journal of basic Engineering* **82** (1), 35–45.

Kitagawa, G. 1996 Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics* **5** (1), 1.

Kitagawa, G. 1998 A Self-Organizing State-Space Model. *Journal of the American Statistical Association* **93** (443), 1203.

Kokkala, J. & Sarkka, S. 2015 On the (non-)convergence of particle filters with Gaussian importance distributions. *IFAC-PapersOnLine* **48** (28), 793–798.

Kong, A., Liu, J. & Wong, W. 1994 Sequencial Imputations and Bayesian Missing Data Problems. *Journal of the American Statistical Association* **89** (425), 278–288.

Liu, J. 1996*a* Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing* .

Liu, J. 2002 *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.

Liu, J. & West, M. 2001 Combined Parameter and State Estimation in Simulation-based Filtering. In *Sequential Monte Carlo Methods in Practice*, pp. 197–223. Springer.

Liu, J. S. 1996*b* Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing* **6** (2), 113–119.

Liu, J. S. & Chen, R. 1998 Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association* **93** (443), 1032–1044.

Lopes, H. F. & Carvalho, C. M. 2011 An Illustrative Introduction to Particle Methods. In *Preliminary Version*, pp. 1–47. None.

Maybeck, P. S. 1979 *Stochastic models, estimation, and control*, , vol. 1.

Moulines, E. 2004 Inference in Hidden Markov Models. *Proceedings of EUSFLAT Conference* .

Murray, L. M., Lee, A. & Jacob, P. E. 2016 Parallel Resampling in the Particle Filter. *Journal of Computational and Graphical Statistics* **25** (3), 789–805.

Nemeth, C., Fearnhead, P. & Mihaylova, L. 2014 Sequential Monte Carlo Methods for State and Parameter Estimation in Abruptly Changing Environments. *IEEE Transactions on Signal Processing* **62** (5), 1245–1255.

Pitt, M. K. & Shephard, N. 1999 Filtering via Simulation: Auxiliary Particle Filters. *Journal of the American Statistical Association* **94** (446), 590–599.

Pitt, M. K., Silva, R. d. S., Giordani, P. & Kohn, R. 2012 On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. In *Journal of Econometrics*, , vol. 171, pp. 134–151.

Prado, R. & West, M. 2010 *Time series: modeling, computation, and inference*. CRC Press.

Press, W. H. & Farrar, G. R. 2012 Recursive Stratified Sampling for Multidimensional Monte Carlo Integration. *Computers in Physics* **190** (1990), 190–195.

Rauch, H. E., Striebel, C. T. & Tung, F. 1965 Maximum likelihood estimates of linear dynamic systems. *AIAA Journal* **3** (8), 1445–1450.

Roberts, G. & Rosenthal, J. 2001 Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science* .

Sherlock, C., Thiery, A. H., Roberts, G. O. & Rosenthal, J. S. 2015 On the efficiency of pseudo-marginal random walk metropolis algorithms. *Annals of Statistics* **43** (1), 238–275.

Shumway, R. H. & Stoffer, D. S. 1982 An approach to time series smoothing and forecasting using the EM algorithm. *Journal of time series analysis* **3** (4), 253–264.

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

Tokdar, S. T. & Kass, R. E. 2010 Importance sampling: A review. *Wiley Interdisciplinary Reviews: Computational Statistics* **2** (1), 54–60.

Van Der Merwe, R., Doucet, a., De Freitas, N. & Wan, E. 2001 The Unscented Particle Filter. *Advances in Neural Information Processing Systems* **96** (6080), 584–590.

Wang, L., Libert, G. & Manneback, P. 1992 Kalman filter algorithm based on singular value decomposition. In *Proceedings of the 31st Conference on Decision and Control*, pp. 1224–1229. IEEE.

West, M. 1993 Approximating Posterior Distributions by Mixture. *Journal of the Royal Statistical Society. Series B (Methodological)* **55** (2), 409–422.

West, M. & Harrison, J. 1997 Bayesian Forecasting and Dynamic Models. *Technometrics* **1** (1), XIV, 682.

West, M., Harrison, P. & Migon, H. 1985 Dynamic generalized linear models and bayesian forecasting. *J. Amer. Statist. Assoc.* **80** (389), 73–97.

Wu, C. 1983 On the Convergence Properties of the {EM} Algorithm. *Annals of Statistics* **11**, 95–103.

Appendix A

PMMH results

A.1 NDLM example

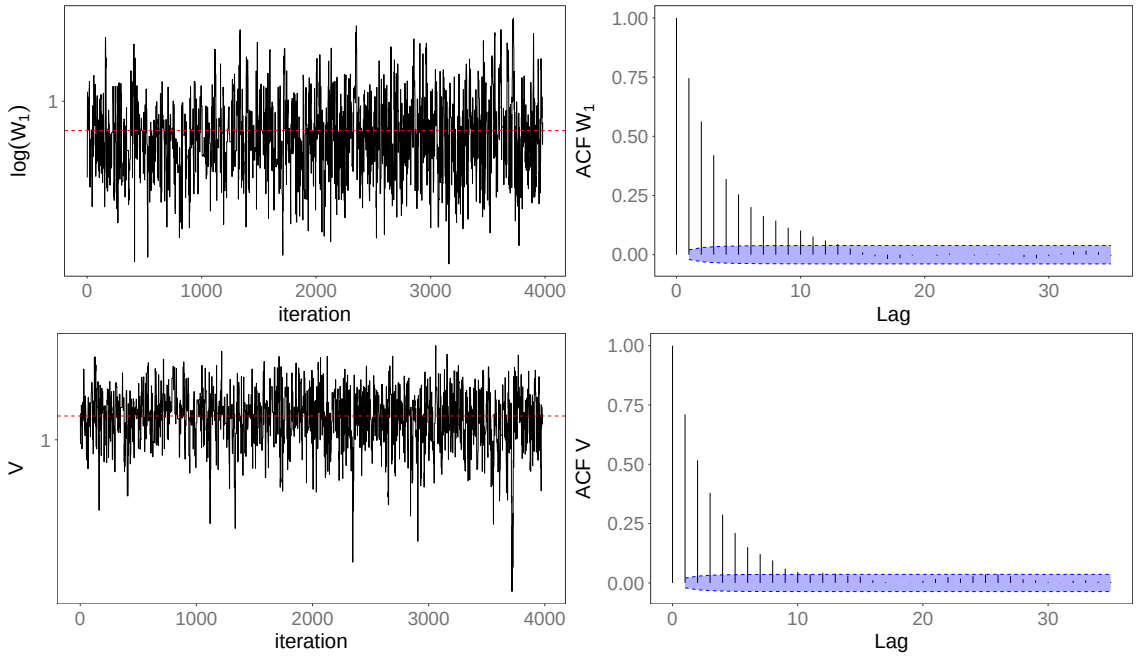


Figure A.1: PMMH traces (*left*) and ACF plots (*right*) for $\boldsymbol{\Phi} = \{\tau^2, \nu^2\}$ for the example Normal DLM in Chapters 8,9 and 10.

A.2 PoDLM example

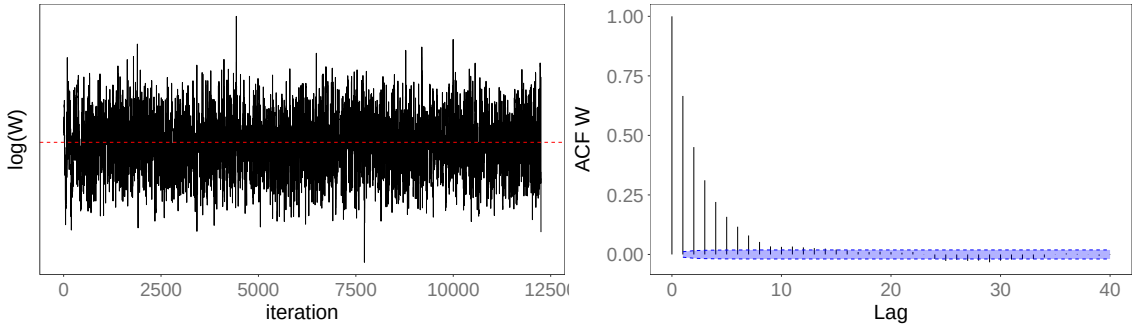


Figure A.2: PMMH traces (*left*) and ACF plots (*right*) for $\boldsymbol{\Phi} = \{\tau^2\}$ for the example Poisson DLM in Chapter 10.

A.3 Particles number

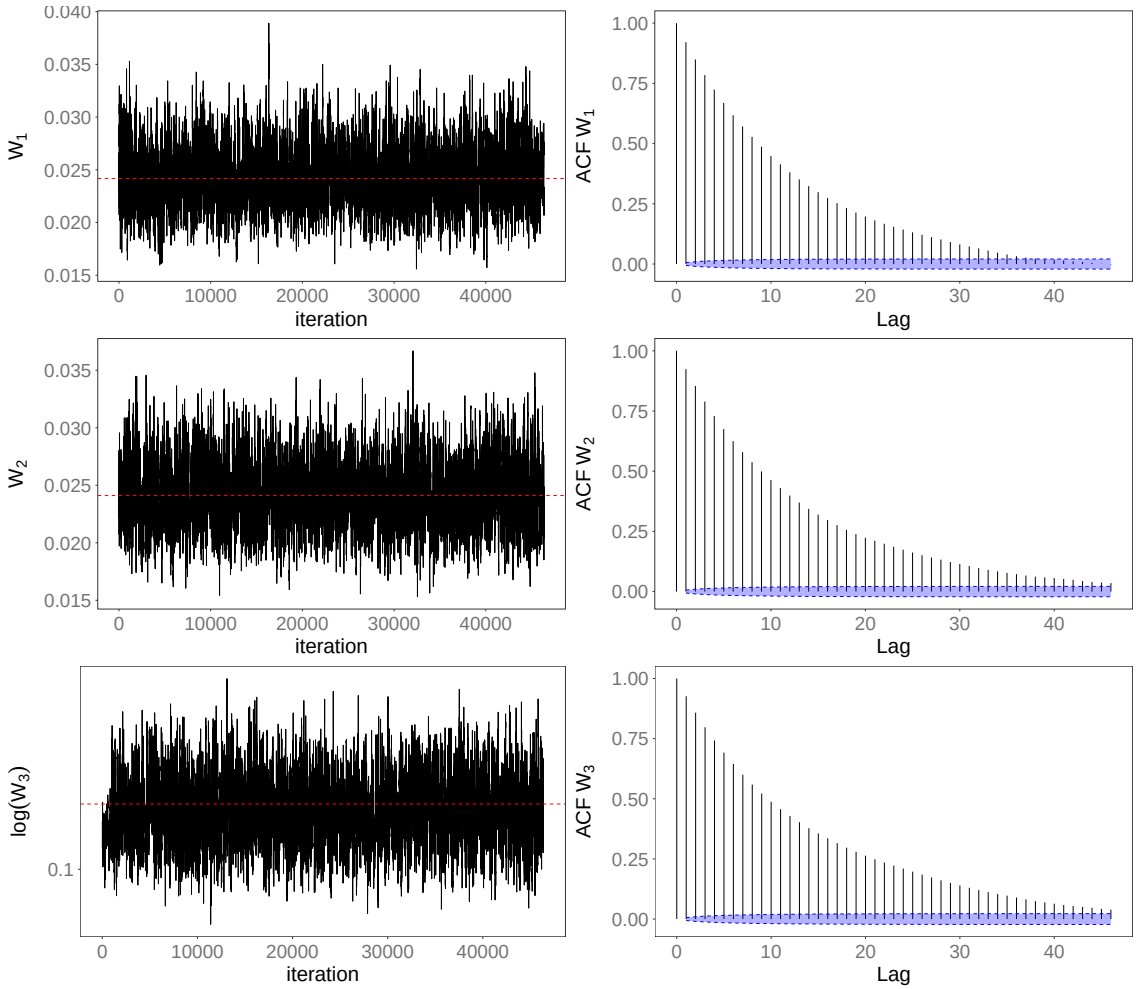


Figure A.3: PMMH traces (*left*) and ACF plots (*right*) for the Poisson DLM in Section 16.1.

A.4 Resampling algorithms

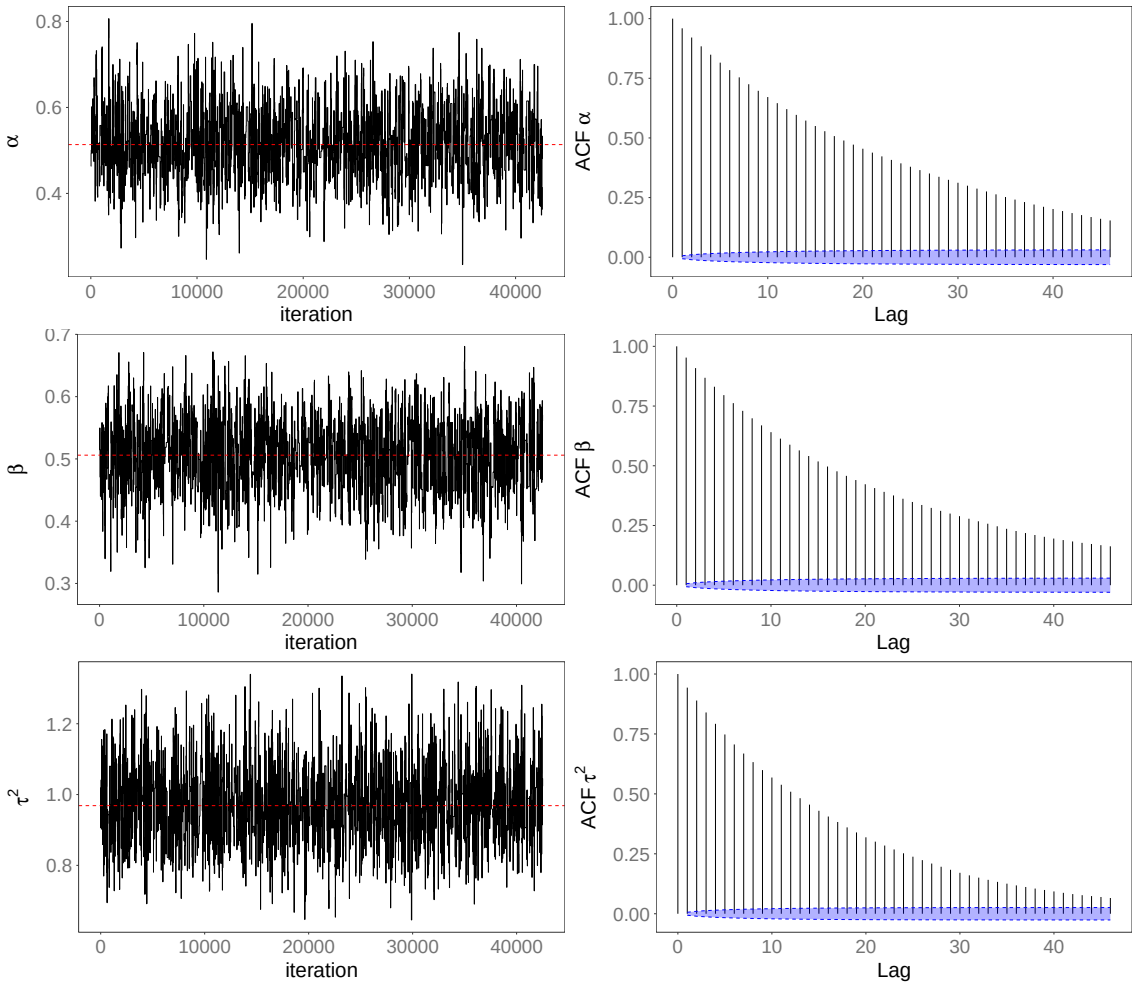


Figure A.4: PMMH traces (*left*) and ACF plots (*right*) for the Poisson AR(1) DLM in Section 16.2.

A.5 Temperature data

A.5.1 Dataset A

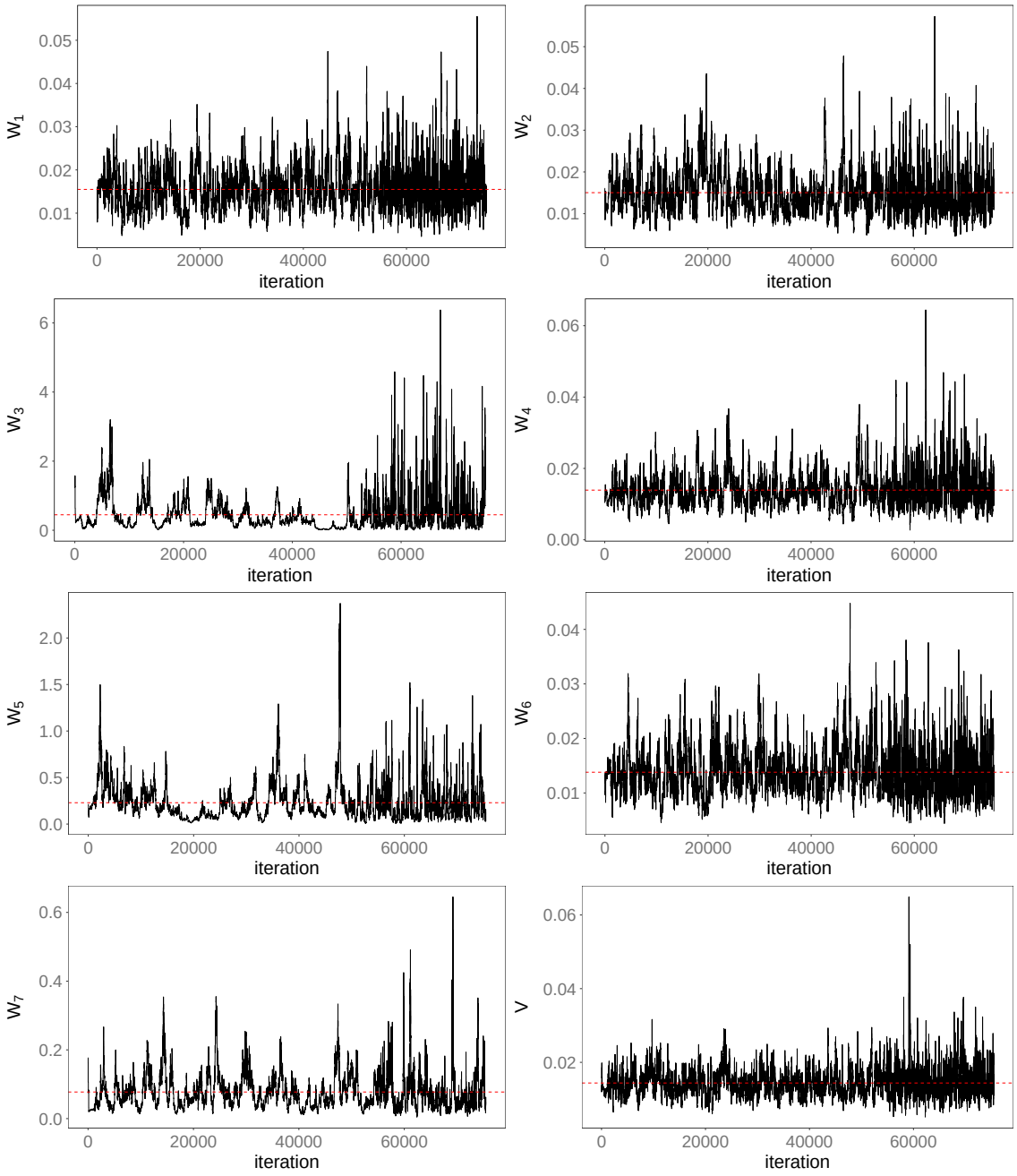


Figure A.5: Temperature dataset A PMMH traces

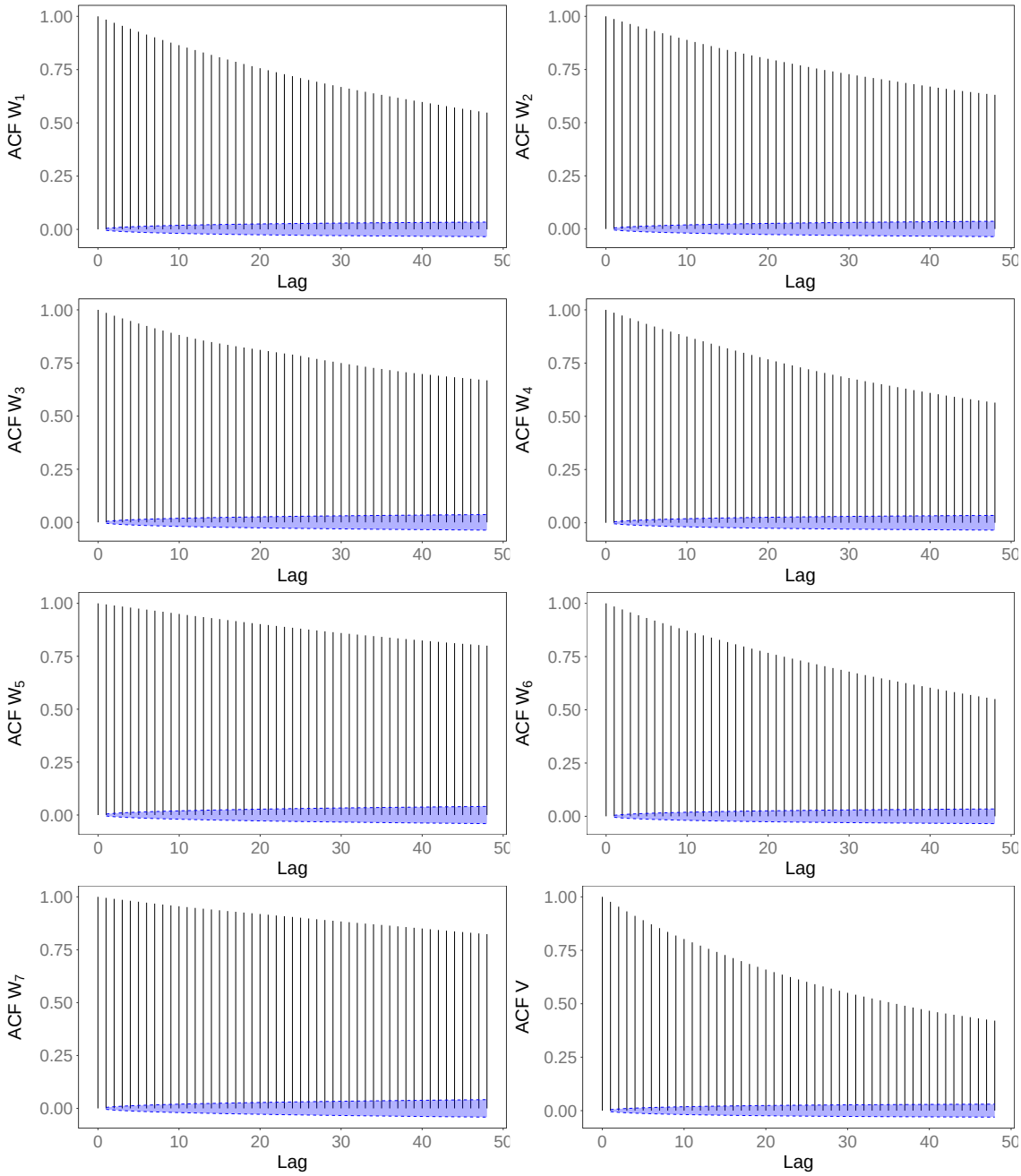


Figure A.6: Temperature data B k-lag ACF

A.5.2 Dataset B

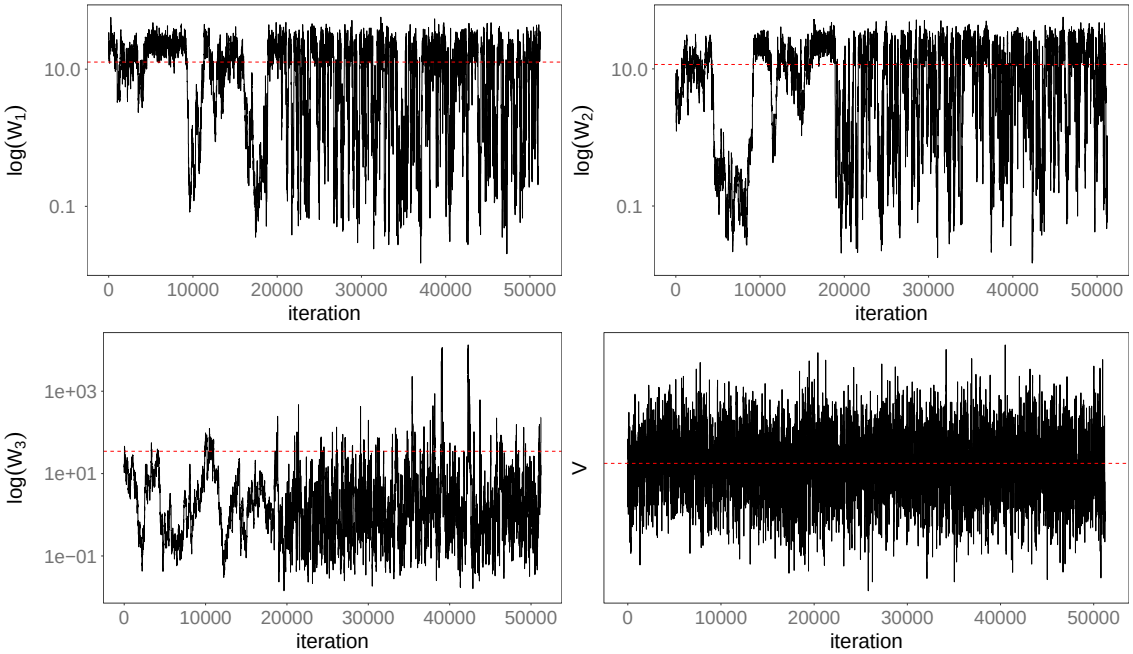


Figure A.7: Temperature dataset B PMMH traces (*log scale*). Horizontal red line is the posterior mean.

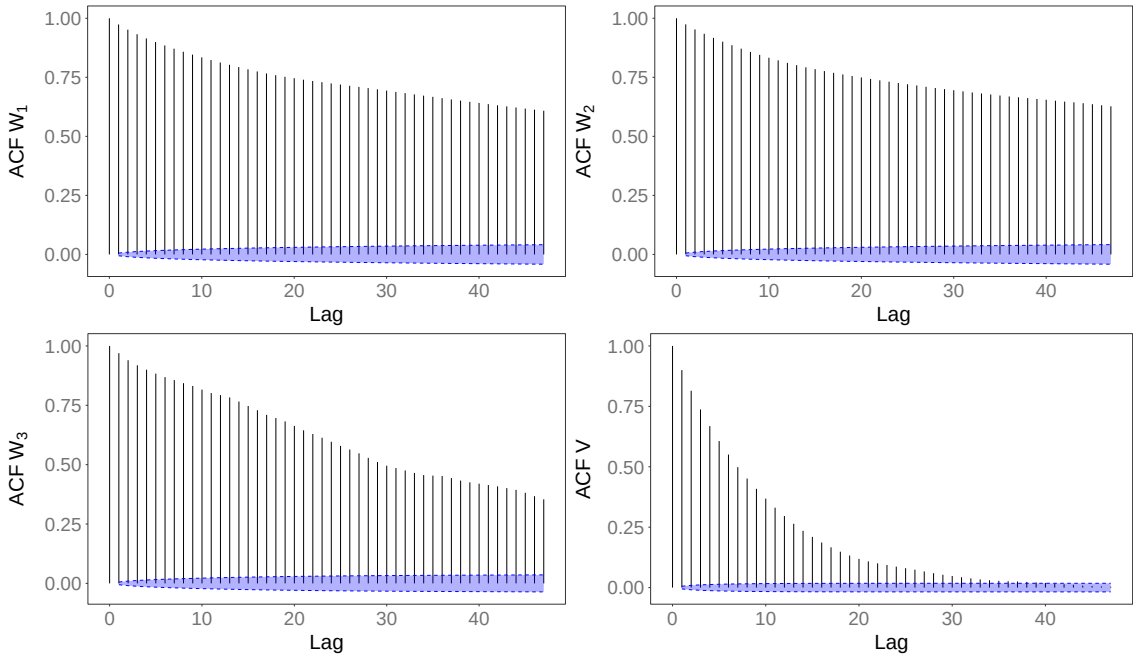


Figure A.8: Temperature dataset B k-lag ACF

A.6 WC98 dataset

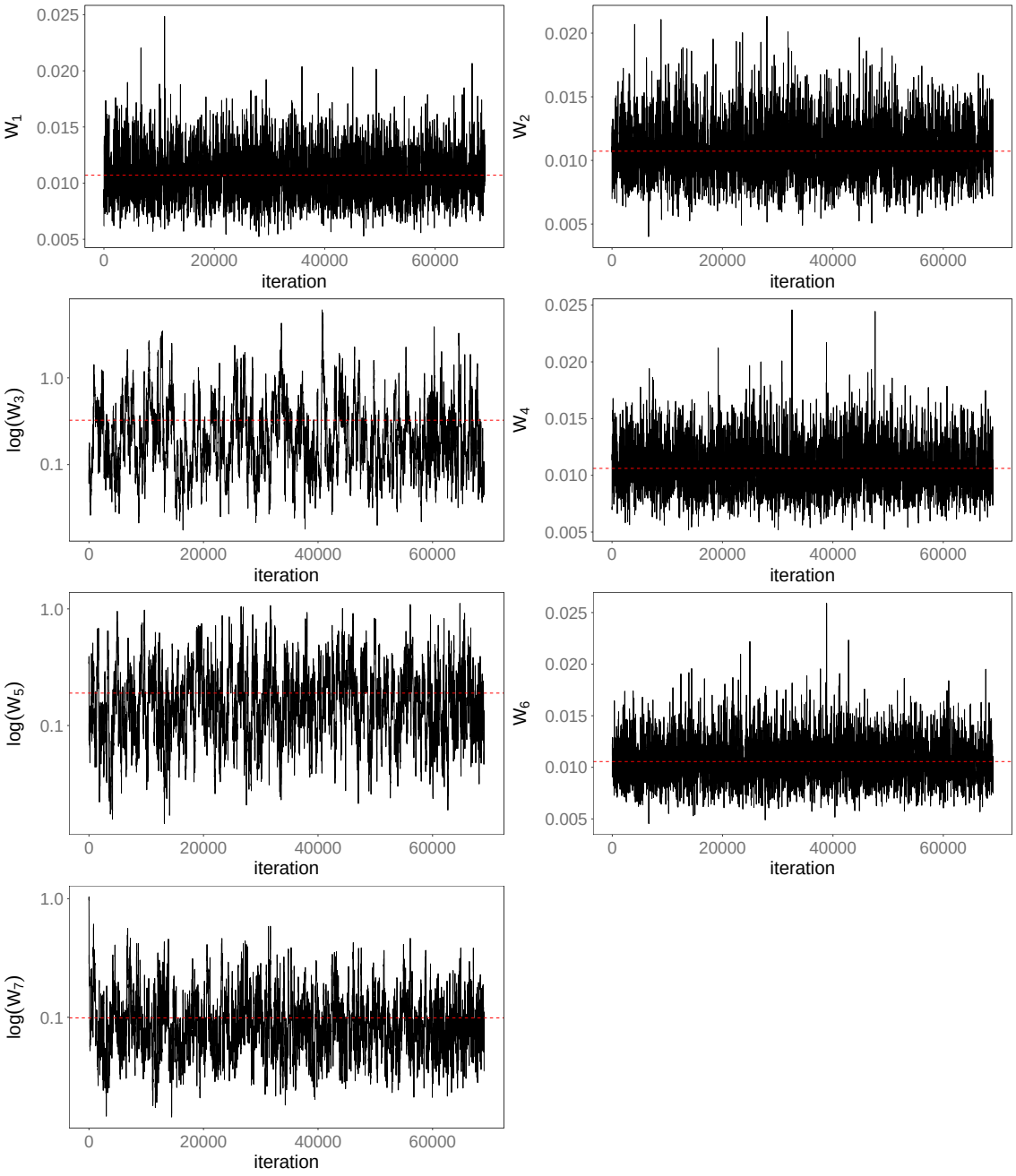


Figure A.9: WC98 PMMH traces

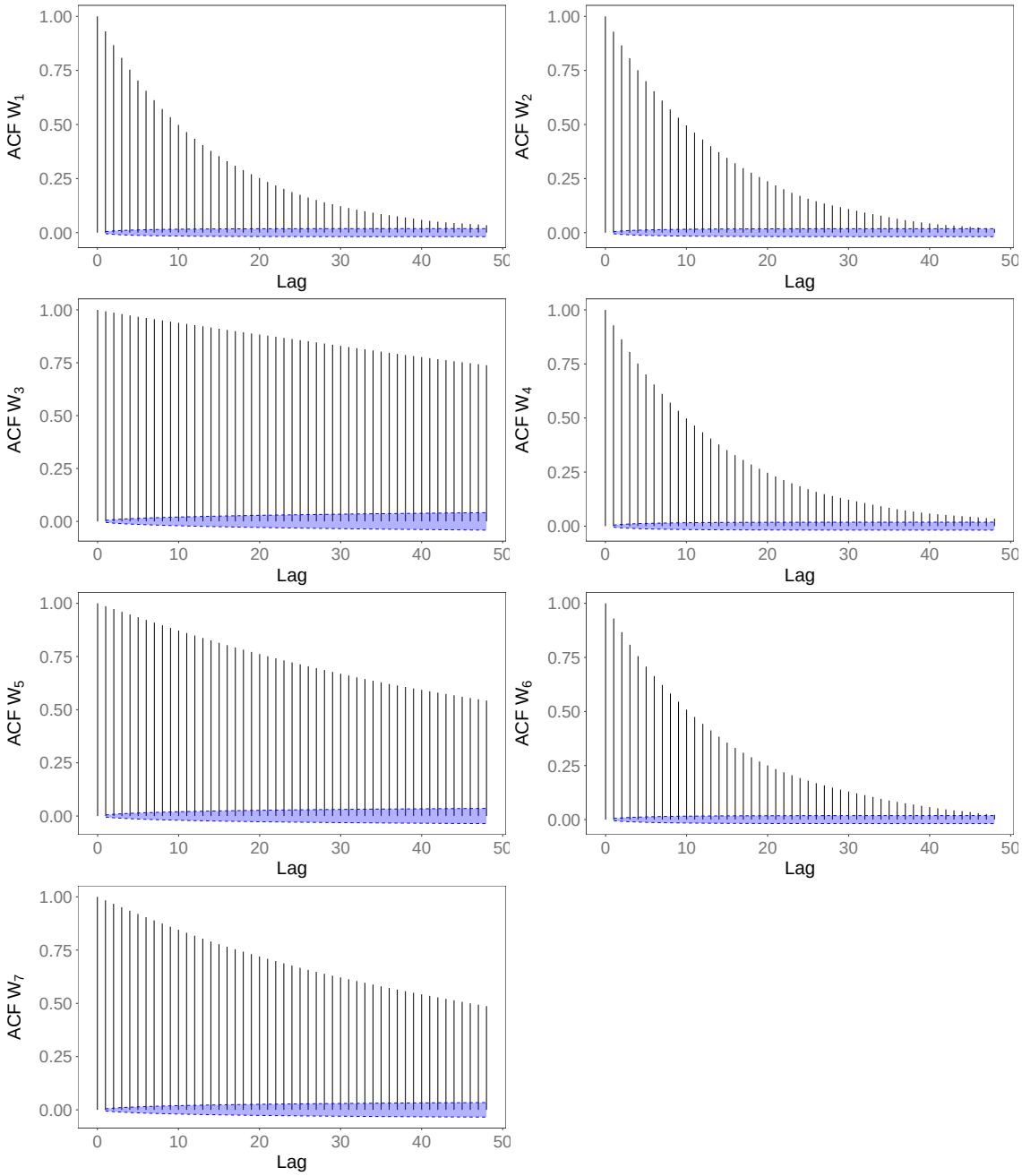


Figure A.10: WC98 k-lag ACF

A.7 Airport data

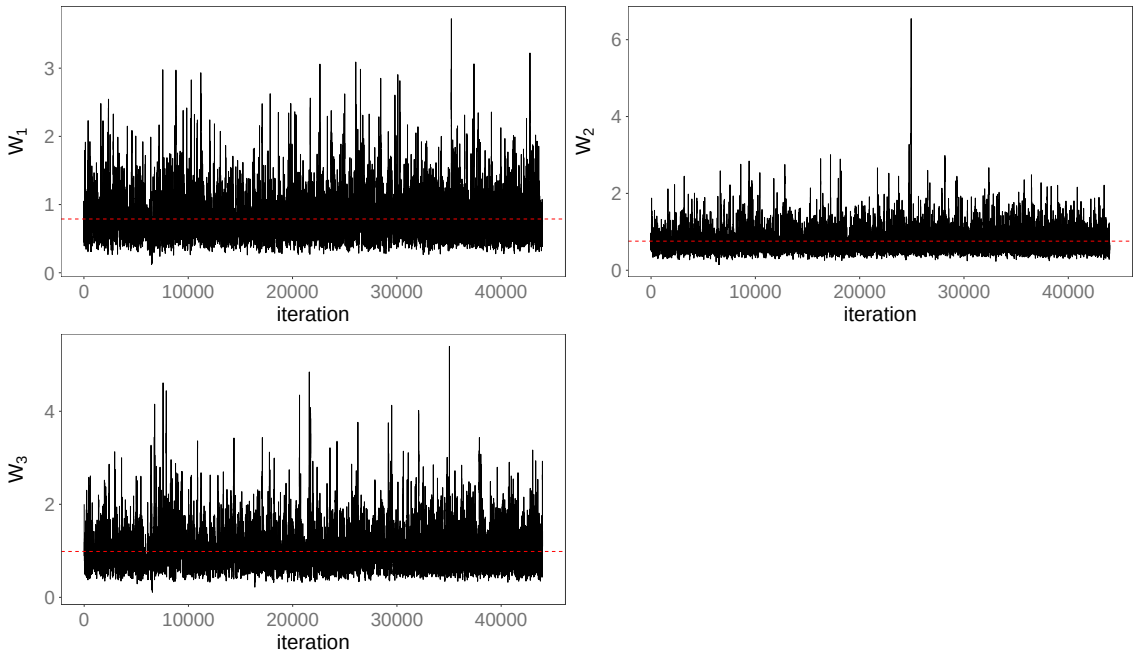


Figure A.11: Airport dataset PMMH traces

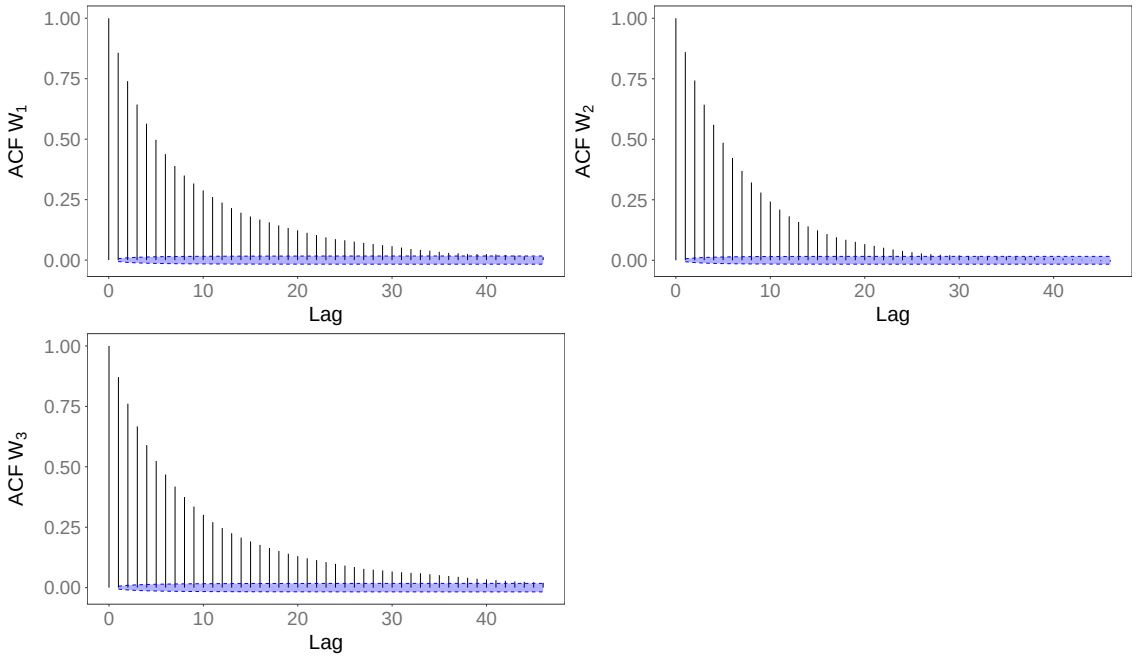
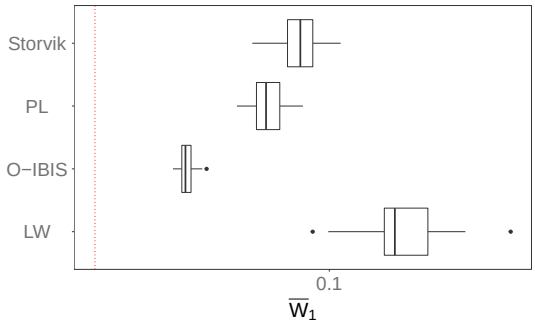


Figure A.12: Airport dataset k-lag ACF

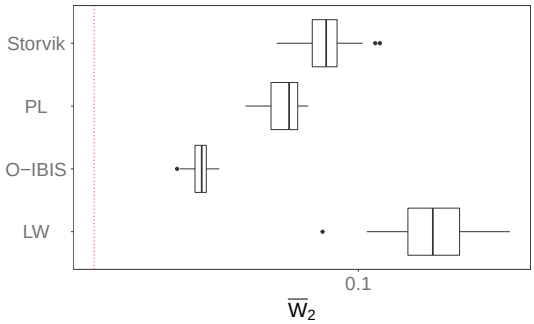
Appendix B

Estimation variability

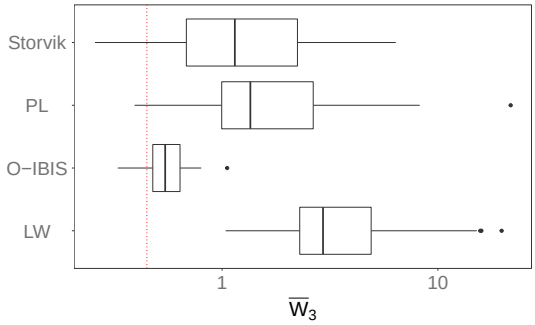
B.1 Temperature data A



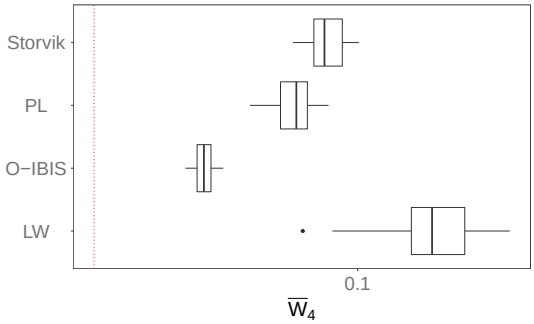
(a) $\overline{W}_1$ posterior means for $n = 50$ runs.



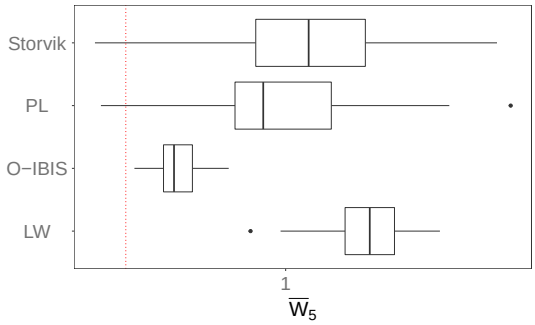
(b) $\overline{W}_2$ posterior means for $n = 50$ runs.



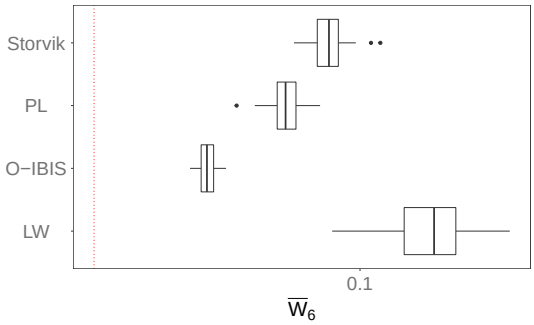
(c) $\overline{W}_3$ posterior means for $n = 50$ runs.



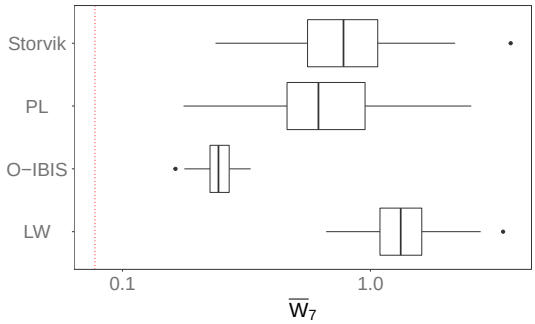
(d) $\overline{W}_4$ posterior means for $n = 50$ runs.



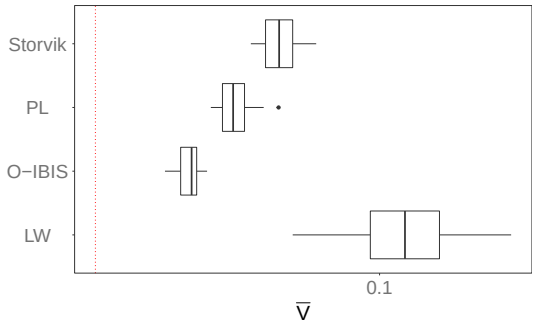
(e) $\overline{W}_5$ posterior means for $n = 50$ runs.



(f) $\overline{W}_6$ posterior means for $n = 50$ runs.

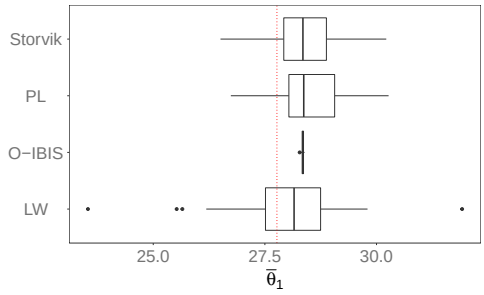


(g) $\overline{W}_7$ posterior means for $n = 50$ runs.

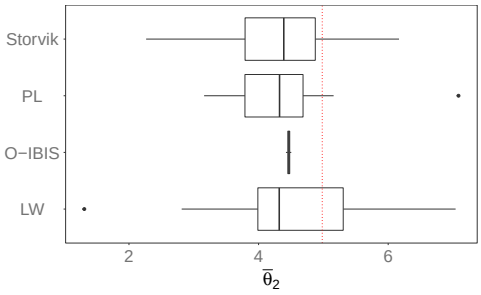


(h) $\overline{V}$ posterior means for $n = 50$ runs.

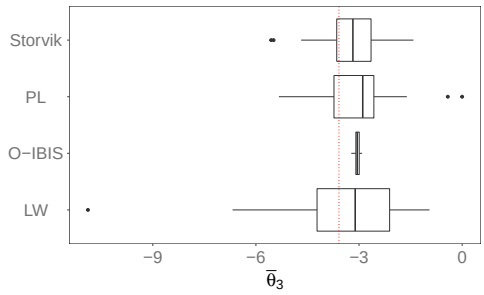
Figure B.1: Parameter posterior means (at $t = N_{obs}$) for $n = 50$ runs with different online filters for the temperature dataset A. Vertical dashed line represents the PMMH posterior mean.



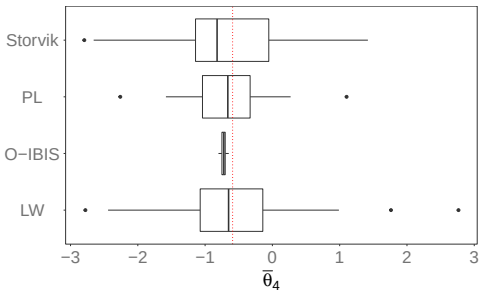
(a) $\overline{\theta}_1$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



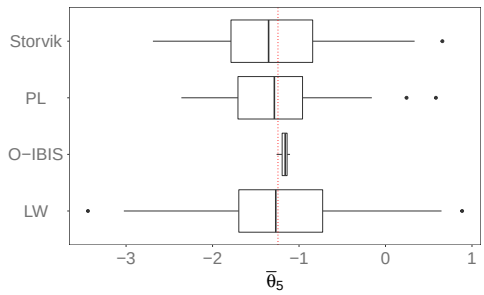
(b) $\overline{\theta}_2$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



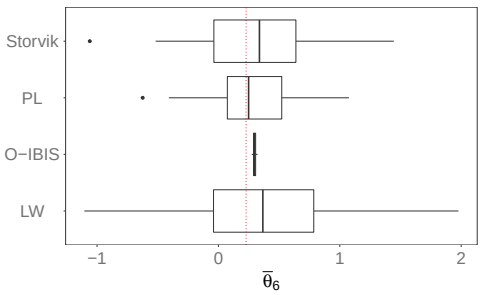
(c) $\overline{\theta}_3$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



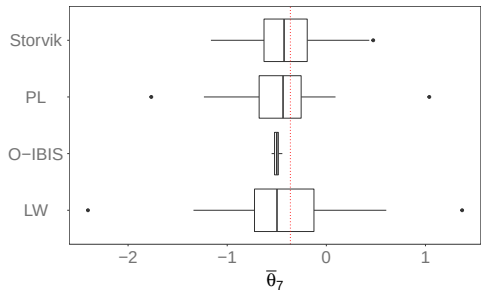
(d) $\overline{\theta}_4$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



(e) $\overline{\theta}_5$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



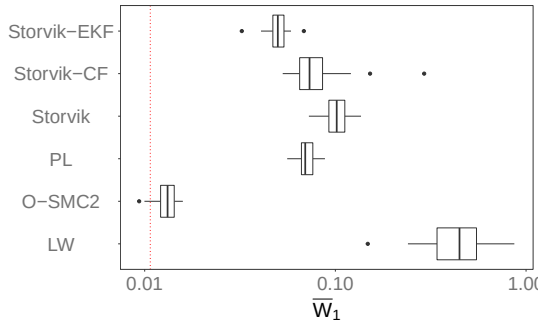
(f) $\overline{\theta}_6$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



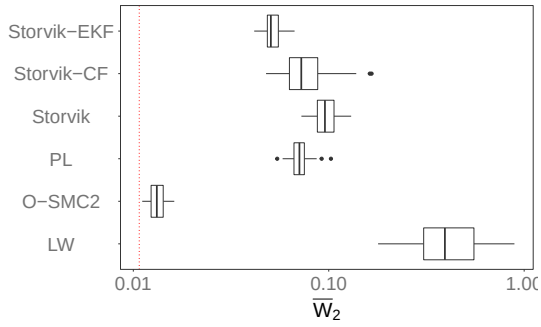
(g) $\overline{\theta}_7$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.

Figure B.2: State posterior means (at $t = N_{obs}$) for $n = 50$ runs with the online filters for the temperature dataset A. Vertical dashed line represents the PMMH state posterior mean..

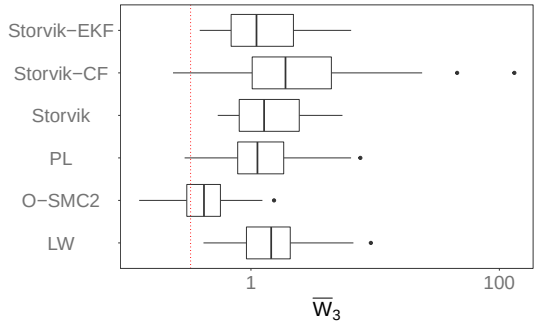
B.2 WC98 data



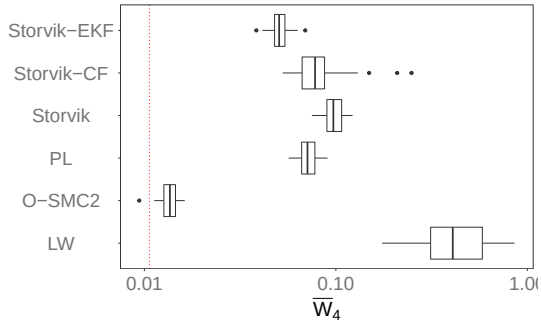
(a) $\overline{W}_1$ posterior means for $n = 50$ runs.



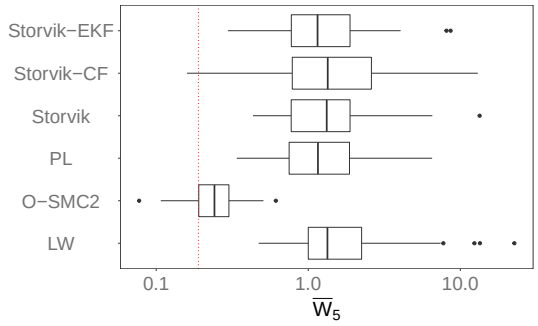
(b) $\overline{W}_2$ posterior means for $n = 50$ runs.



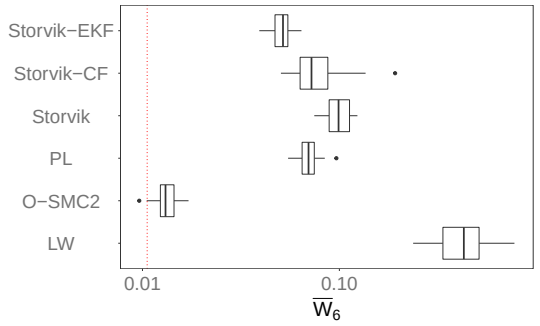
(c) $\overline{W}_3$ posterior means for $n = 50$ runs.



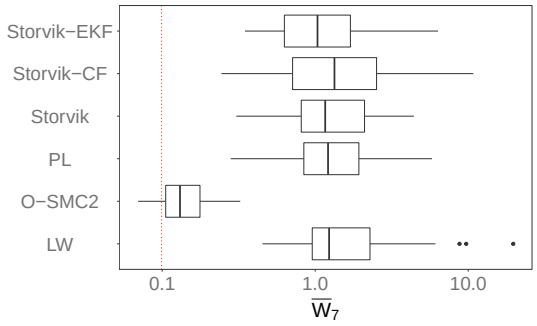
(d) $\overline{W}_4$ posterior means for $n = 50$ runs.



(e) $\overline{W}_5$ posterior means for $n = 50$ runs.

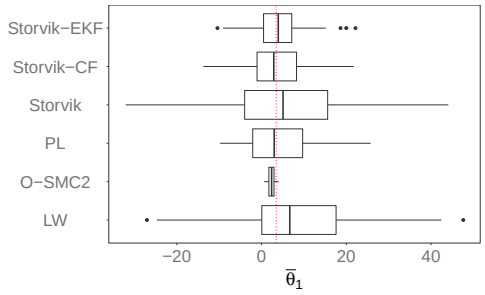


(f) $\overline{W}_6$ posterior means for $n = 50$ runs.

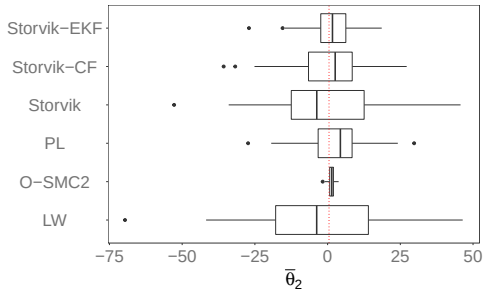


(g) $\overline{W}_7$ posterior means for $n = 50$ runs.

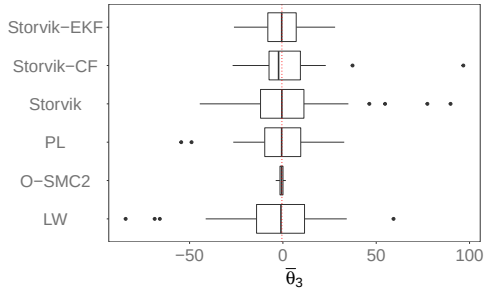
Figure B.3: Parameter posterior means (at $t = N_{obs}$) for $n = 50$ runs with different online filters for the WC98 dataset. Vertical dashed line represents the PMMH posterior mean.



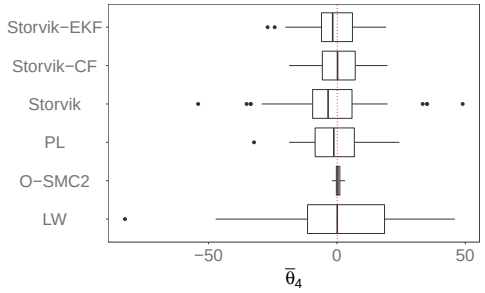
(a) $\overline{\boldsymbol{\theta}}_1$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



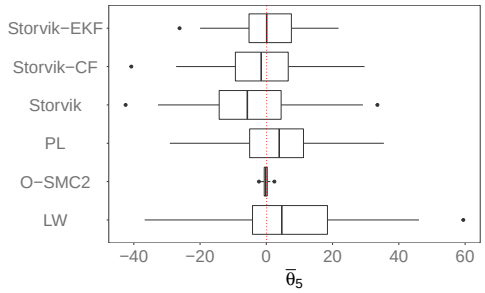
(b) $\overline{\boldsymbol{\theta}}_2$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



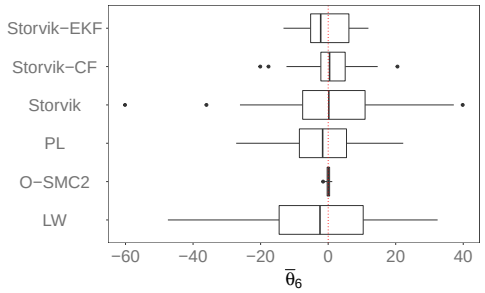
(c) $\overline{\boldsymbol{\theta}}_3$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



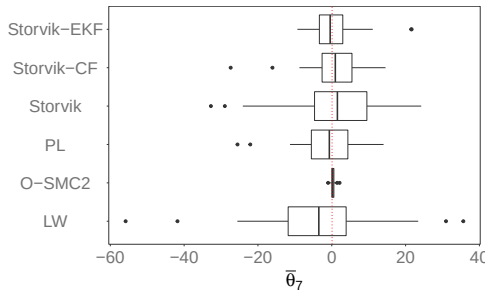
(d) $\overline{\boldsymbol{\theta}}_4$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



(e) $\overline{\boldsymbol{\theta}}_5$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.

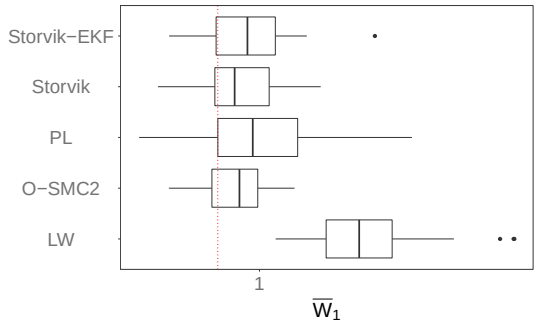


(f) $\overline{\boldsymbol{\theta}}_6$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.

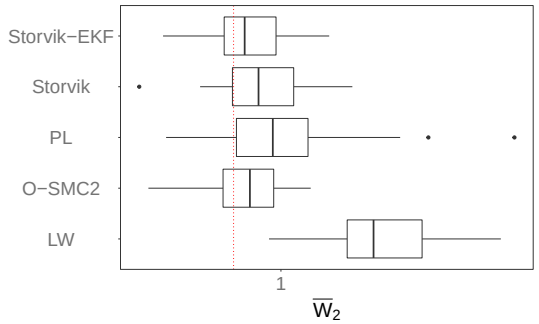


(g) $\overline{\boldsymbol{\theta}}_7$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.

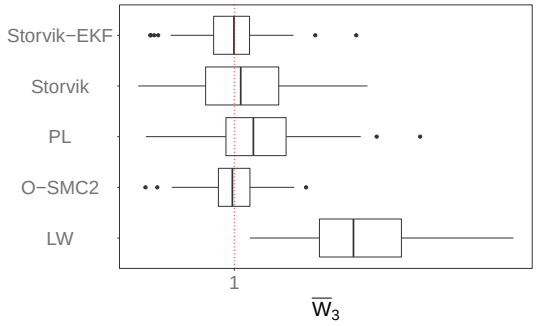
Figure B.4: State posterior means (at $t = N_{obs}$) for $n = 50$ runs with the online filters for the WC98 dataset. Vertical dashed line represents the PMMH state posterior mean..



(a) $\overline{W}_1$ posterior means for $n = 50$ runs.



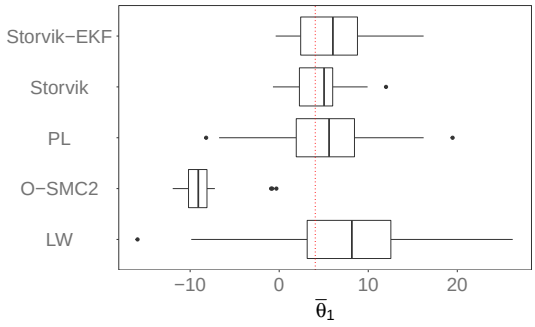
(b) $\overline{W}_2$ posterior means for $n = 50$ runs.



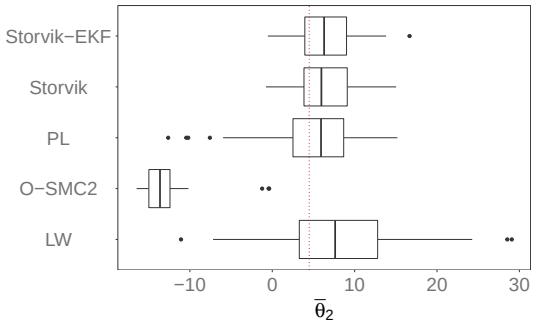
(c) $\overline{W}_3$ posterior means for $n = 50$ runs.

Figure B.5: Parameter posterior means (at $t = N_{obs}$) for $n = 50$ runs with different online filters for the airport dataset. Vertical dashed line represents the PMMH posterior mean.

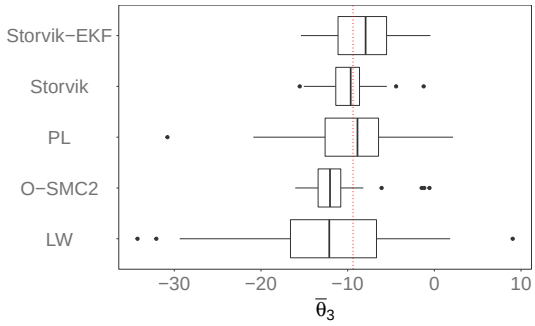
B.3 Airport data



(a) $\overline{\theta}_1$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



(b) $\overline{\theta}_2$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



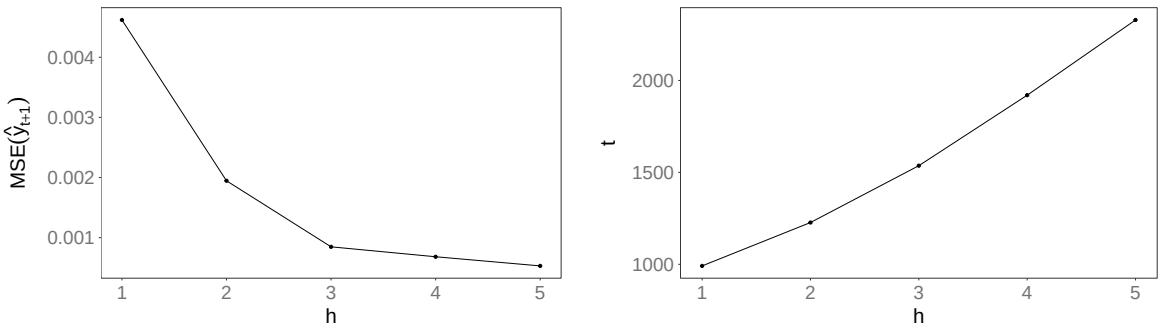
(c) $\overline{\theta}_3$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.

Figure B.6: State posterior means (at $t = N_{obs}$) for $n = 50$ runs with the online filters for the airport dataset. Vertical dashed line represents the PMMH state posterior mean..

Appendix C

Number of harmonics

C.1 Temperature dataset A



(a) MSE between y_{t+1} and $\widehat{y}_{t+1}$.

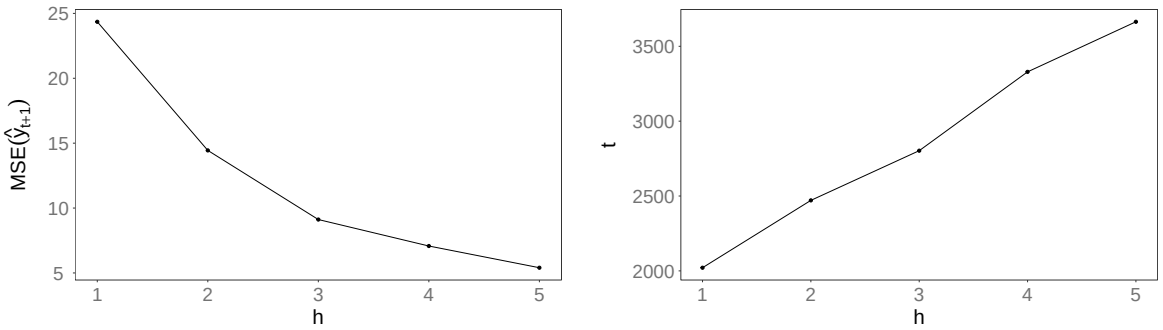
(b) Computational cost.

Figure C.1: MSE (*left*) and computational cost (*right*) using Storvik on the temperature dataset A for different numbers of harmonics..

Harmonics	MSE	time (s)
1	0.00461	991.409
2	0.00194	1227.318
3	0.00084	1536.417
4	0.00068	1919.874
5	0.00053	2329.045

Table C.1: Summary of one-step ahead observation forecast MSE and computational cost for a varying number of harmonics using Storvik on the temperature dataset A.

C.2 WC98



(a) MSE between y_{t+1} and $\widehat{y}_{t+1}$.

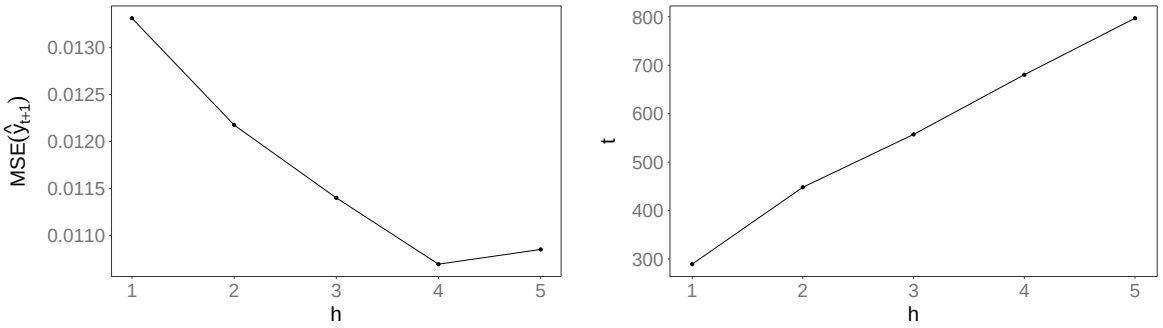
(b) Computational cost.

Figure C.2: MSE (*left*) and computational cost (*right*) using Storvik on the WC98 dataset for different numbers of harmonics..

Harmonics	MSE	time (s)
1	24.3538	2021.080
2	14.4420	2470.846
3	9.1153	2802.747
4	7.0754	3329.307
5	5.4026	3663.765

Table C.2: Summary of one-step ahead observation forecast MSE and computational cost for a varying number of harmonics using Storvik on the WC98 dataset.

C.3 Airport data



(a) MSE between y_{t+1} and $\widehat{y}_{t+1}$.

(b) Computational cost.

Figure C.3: MSE (*left*) and computational cost (*right*) using Storvik on the airport dataset for different numbers of harmonics..

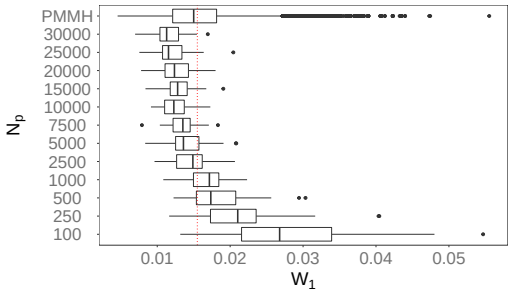
Harmonics	MSE	time (s)
1	0.0133	289.40
2	0.0121	448.13
3	0.0114	557.12
4	0.0106	680.33
5	0.0108	796.98

Table C.3: Summary of one-step ahead observation forecast MSE and computational cost for a varying number of harmonics using Storvik on the airport dataset.

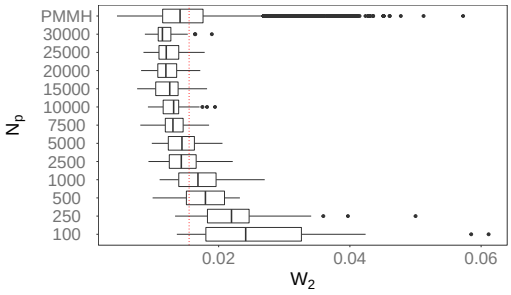
Appendix D

Number of particles

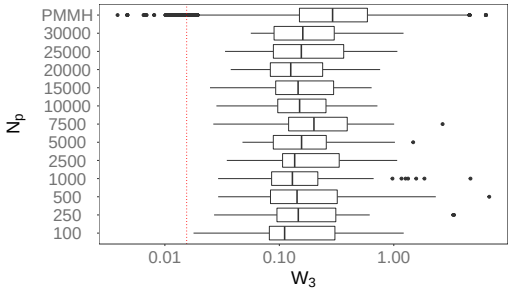
D.1 Temperature dataset A



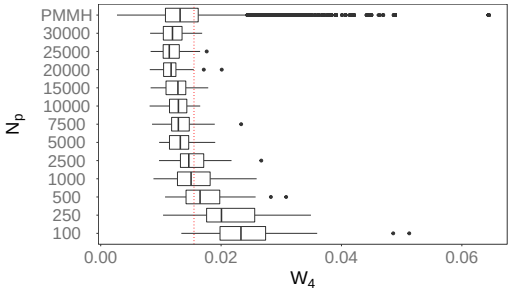
(a) $\overline{W}_1$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



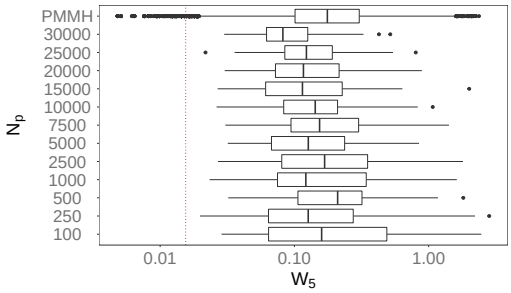
(b) $\overline{W}_2$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



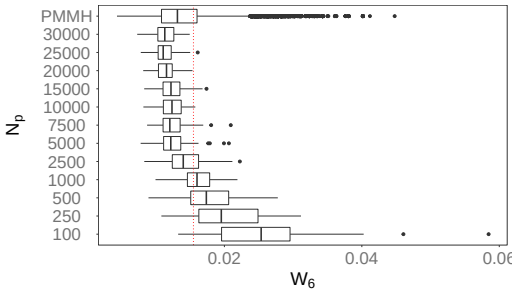
(c) $\overline{W}_3$ posterior means (at $t = N_{obs}$) for $n = 50$ runs (*log-scale*).



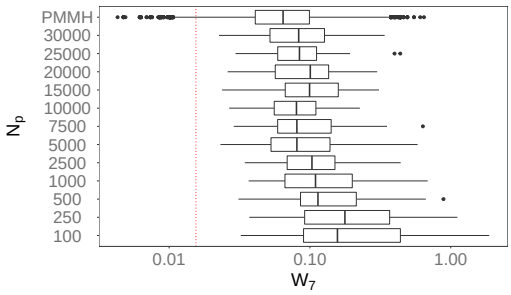
(d) $\overline{W}_4$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



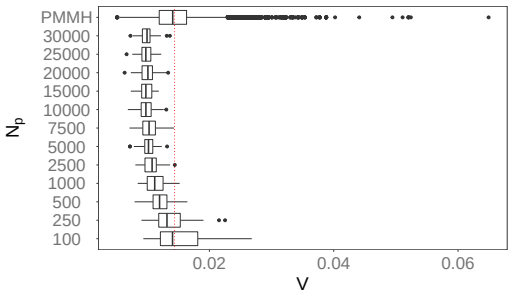
(e) $\overline{W}_5$ posterior means (at $t = N_{obs}$) for $n = 50$ runs (*log-scale*).



(f) $\overline{W}_6$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.

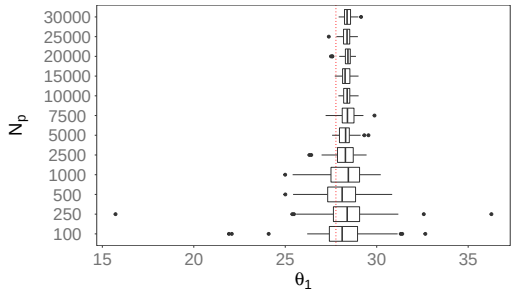


(g) $\overline{W}_7$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.

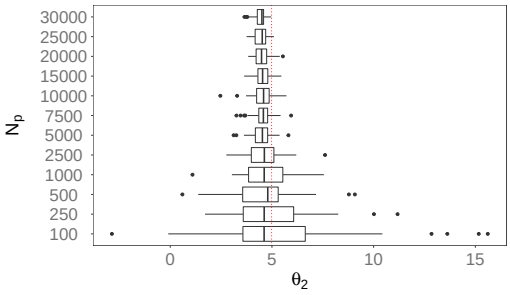


(h) $\overline{V}$ posterior means (at $t = N_{obs}$) for $n = 50$ runs (*log-scale*).

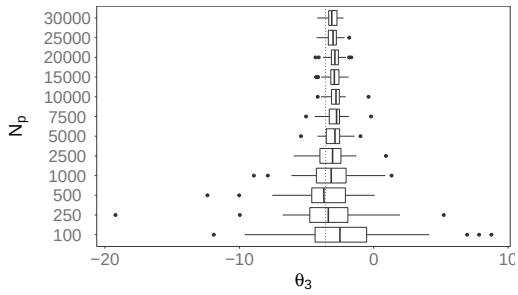
Figure D.1: Parameter posterior means (at $t = N_{obs}$) for $n = 50$ runs with the online filters for the temperature A dataset with varying N_p. Vertical dashed line represents the PMMH parameter posterior mean..



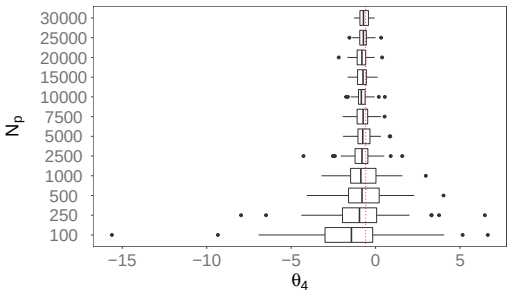
(a) $\overline{\boldsymbol{\theta}}_1$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



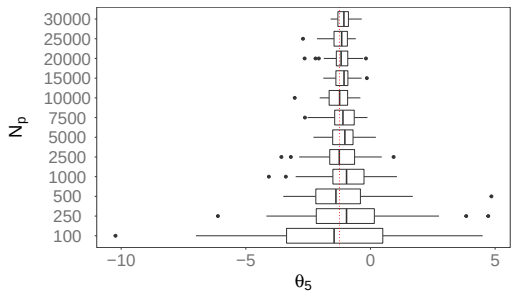
(b) $\overline{\boldsymbol{\theta}}_2$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



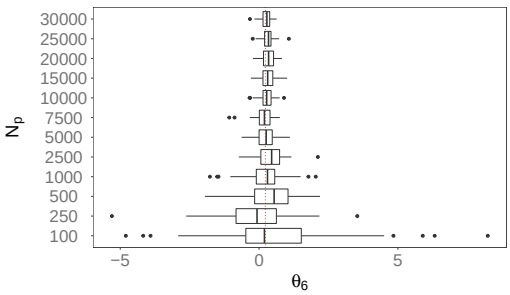
(c) $\overline{\boldsymbol{\theta}}_3$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



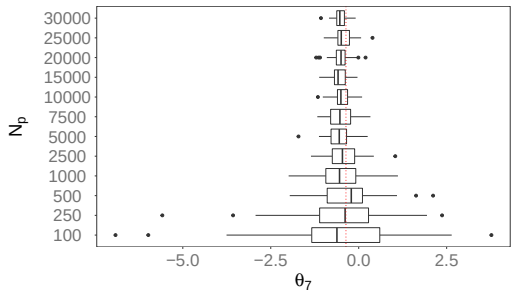
(d) $\overline{\boldsymbol{\theta}}_4$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



(e) $\overline{\boldsymbol{\theta}}_5$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



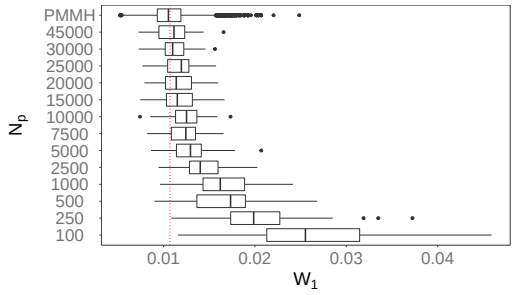
(f) $\overline{\boldsymbol{\theta}}_6$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



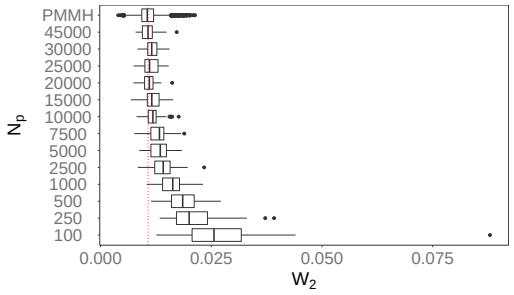
(g) $\overline{\boldsymbol{\theta}}_7$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.

Figure D.2: State posterior means (at $t = N_{obs}$) for $n = 50$ runs with the online filters for the temperature A dataset with varying N_p. Vertical dashed line represents the PMMH state posterior mean..

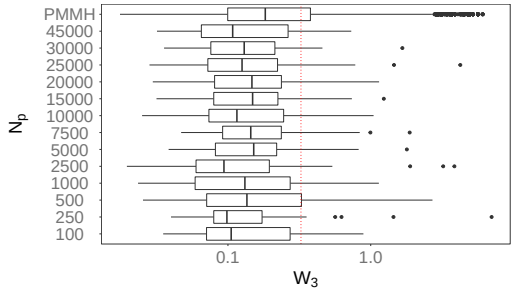
D.2 WC98



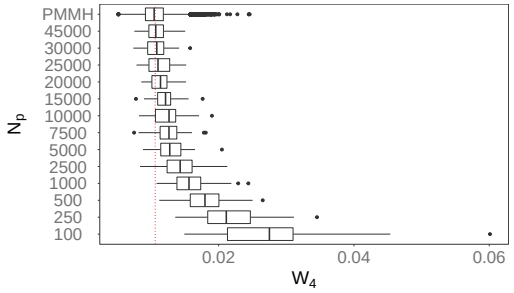
(a) $\overline{W}_1$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



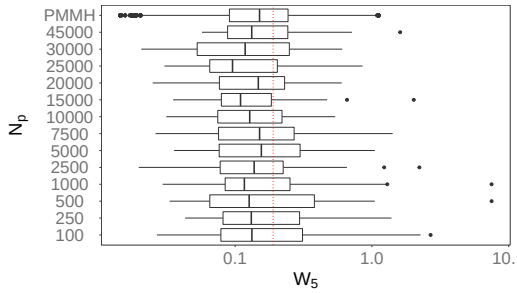
(b) $\overline{W}_2$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



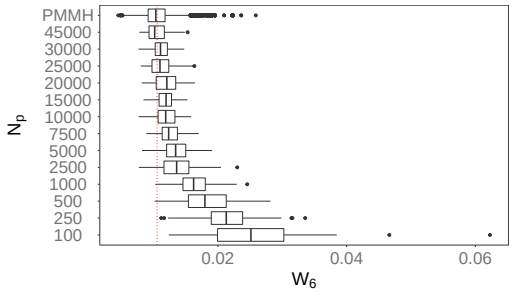
(c) $\overline{W}_3$ posterior means (at $t = N_{obs}$) for $n = 50$ runs (*log=scale*).



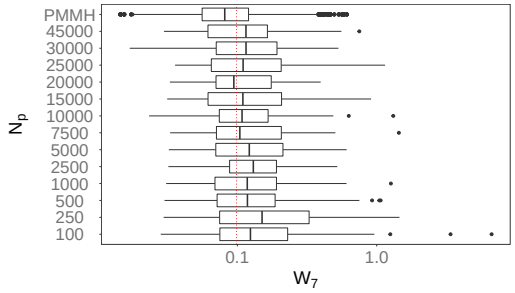
(d) $\overline{W}_4$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



(e) $\overline{W}_5$ posterior means (at $t = N_{obs}$) for $n = 50$ runs (*log=scale*).

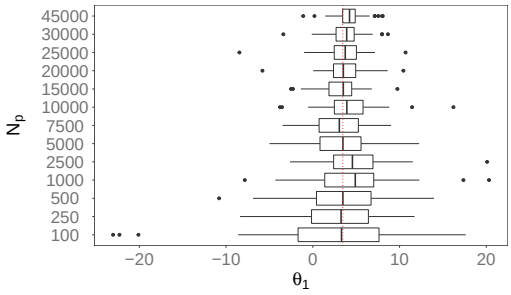


(f) $\overline{W}_6$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.

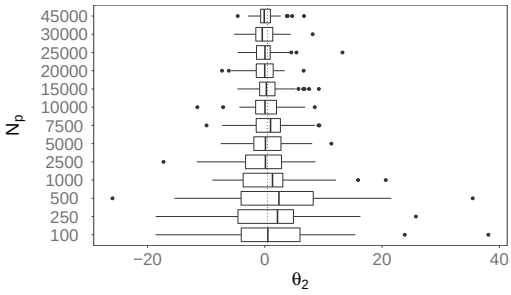


(g) $\overline{W}_7$ posterior means (at $t = N_{obs}$) for $n = 50$ runs (*log=scale*).

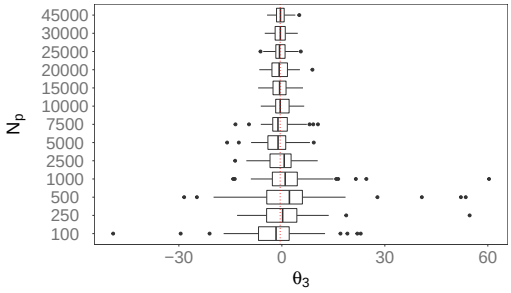
Figure D.3: Parameter posterior means (at $t = N_{obs}$) for $n = 50$ runs with the online filters for the WC98 dataset with varying N_p. Vertical dashed line represents the PMMH state posterior mean..



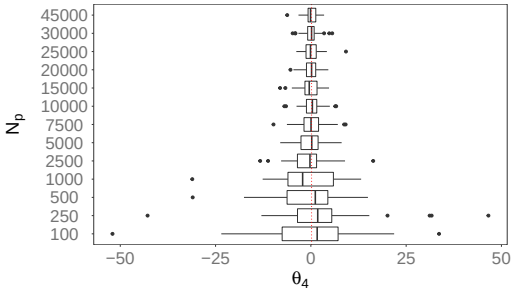
(a) $\overline{\boldsymbol{\theta}}_1$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



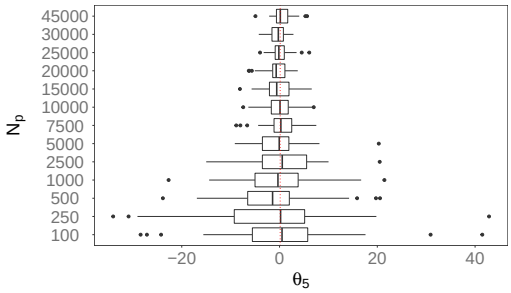
(b) $\overline{\boldsymbol{\theta}}_2$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



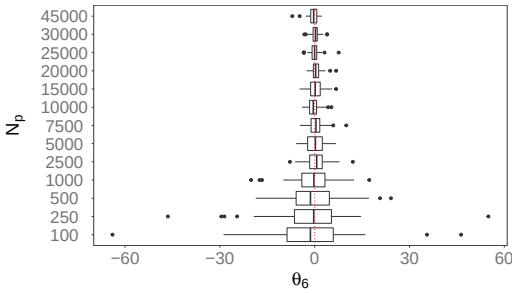
(c) $\overline{\boldsymbol{\theta}}_3$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



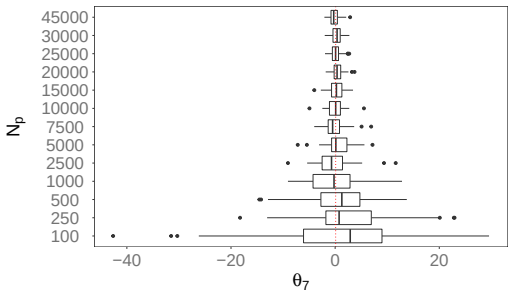
(d) $\overline{\boldsymbol{\theta}}_4$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



(e) $\overline{\boldsymbol{\theta}}_5$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



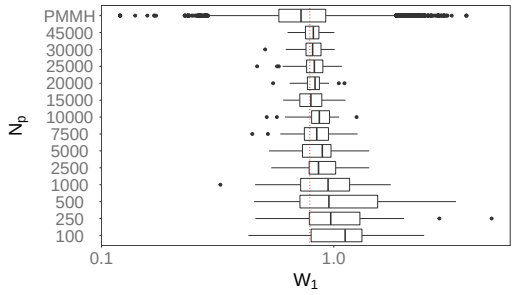
(f) $\overline{\boldsymbol{\theta}}_6$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



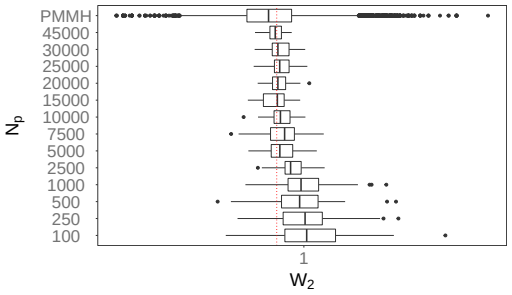
(g) $\overline{\theta}_7$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.

Figure D.4: State posterior means (at $t = N_{obs}$) for $n = 50$ runs with the online filters for the WC98 dataset with varying N_p. Vertical dashed line represents the PMMH state posterior mean..

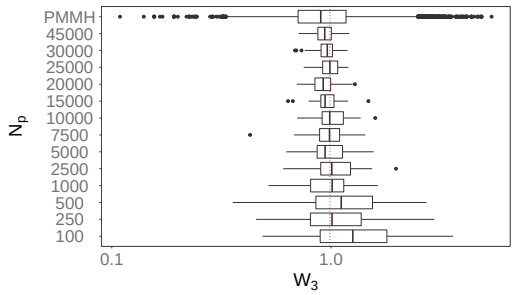
D.3 Airport data



(a) $\overline{W}_1$ posterior means (at $t = N_{obs}$) for $n = 50$ runs (*log-scale*).

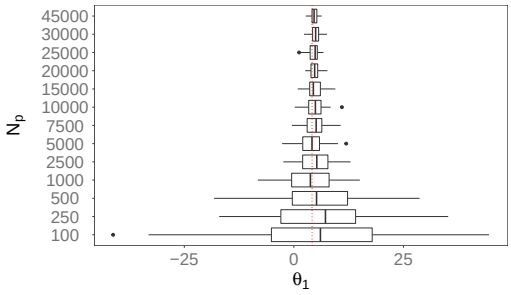


(b) $\overline{W}_2$ posterior means (at $t = N_{obs}$) for $n = 50$ runs (*log-scale*).

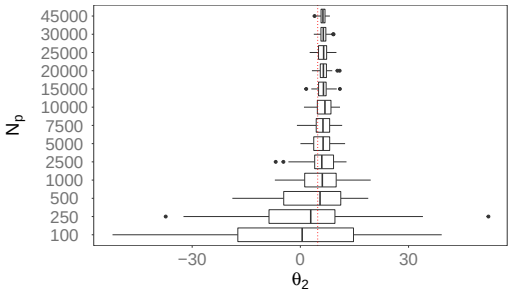


(c) $\overline{W}_3$ posterior means (at $t = N_{obs}$) for $n = 50$ runs (*log-scale*).

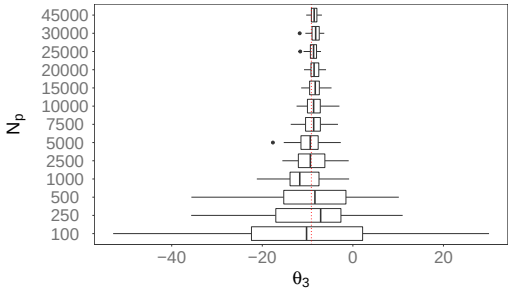
Figure D.5: Parameter posterior means (at $t = N_{obs}$) for $n = 50$ runs with the online filters for the airport dataset with varying N_p. Vertical dashed line represents the PMMH state posterior mean..



(a) $\overline{\boldsymbol{\theta}}_1$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



(b) $\overline{\boldsymbol{\theta}}_2$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.



(c) $\overline{\boldsymbol{\theta}}_3$ posterior means (at $t = N_{obs}$) for $n = 50$ runs.

Figure D.6: State posterior means (at $t = N_{obs}$) for $n = 50$ runs with the online filters for the airport dataset with varying N_p. Vertical dashed line represents the PMMH state posterior mean..

Appendix E

Smoothing parameter

E.1 Temperature data

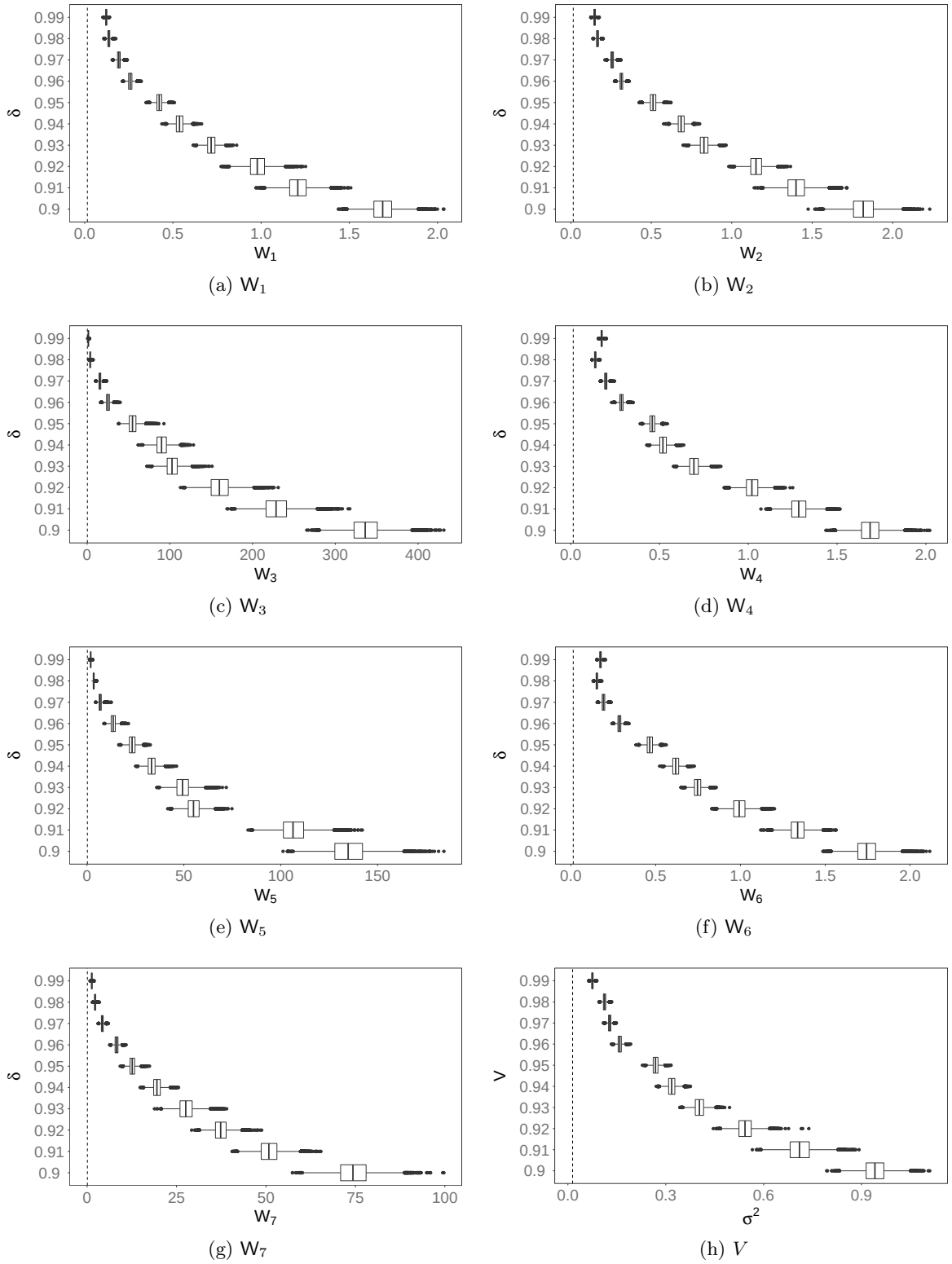
E.1.1 Dataset A

δ	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$	$\overline{\mathsf{W}}_4$	$\overline{\mathsf{W}}_5$	$\overline{\mathsf{W}}_6$	$\overline{\mathsf{W}}_7$	$\overline{V}$
0.9	1.6905	1.82	337.1371	1.6879	135.2982	1.7457	74.5457	0.9417
0.91	1.2094	1.4025	229.2419	1.2855	106.7359	1.338	50.9257	0.711
0.92	0.9803	1.1524	160.7837	1.0219	55.1864	0.9941	37.4144	0.5437
0.93	0.7178	0.8288	103.2251	0.695	49.5772	0.7474	27.7628	0.4036
0.94	0.5384	0.6862	90.4689	0.5204	33.4591	0.6183	19.5803	0.3184
0.95	0.4234	0.5119	55.4202	0.4597	23.4323	0.4658	12.657	0.269
0.96	0.2585	0.3143	25.431	0.2854	13.6447	0.286	8.2404	0.159
0.97	0.194	0.2572	15.6092	0.1972	6.8814	0.1923	4.26	0.1281
0.98	0.1371	0.1661	3.8463	0.138	3.4961	0.1531	2.2309	0.1124
0.99	0.1223	0.1480	1.6887	0.1740	1.9649	0.1746	1.3159	0.0752
PMMH	0.0155	0.0150	0.4481	0.0139	0.2295	0.0138	0.0774	0.0144

Table E.1: Summary of parameter posterior mean estimation with L&W for the temperature dataset A with varying δ

δ	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	$\boldsymbol{\theta}_4$	$\boldsymbol{\theta}_5$	$\boldsymbol{\theta}_6$	$\boldsymbol{\theta}_7$
0.9	10.3175	14.6803	27.7929	9.7076	8.4063	4.9006	4.3048
0.91	8.7966	13.1653	21.6054	11.8184	10.4796	4.1353	4.0284
0.92	6.0817	8.4377	12.6426	5.6472	6.3974	3.9255	3.3768
0.93	1.9293	5.1469	12.7155	5.4308	4.8993	2.3383	2.1724
0.94	1.6119	4.2967	9.8588	4.2224	3.4841	1.6956	1.7079
0.95	5.9177	6.5925	6.3583	3.6793	3.6638	1.5829	1.5177
0.96	3.375	6.6968	12.2035	6.2234	6.8612	4.3876	3.1077
0.97	1.645	3.8503	4.4991	1.9811	1.8922	1.2445	0.7889
0.98	1.7477	3.7198	4.3917	2.0077	1.7707	1.2214	0.5689
0.99	1.5548	5.0378	5.5941	3.0167	2.8366	1.6267	0.7348

Table E.2: Summary of the state posterior mean MSE compared to PMMH using L&W for the temperature dataset A



(a) W_1 (b) W_2

(c) W_3 (d) W_4

(e) W_5 (f) W_6

(g) W_7 (h) V

Figure E.1: Parameter posterior estimation using L&W with different δ for the dataset A. Vertical dashed line represents PMMH estimated posterior mean

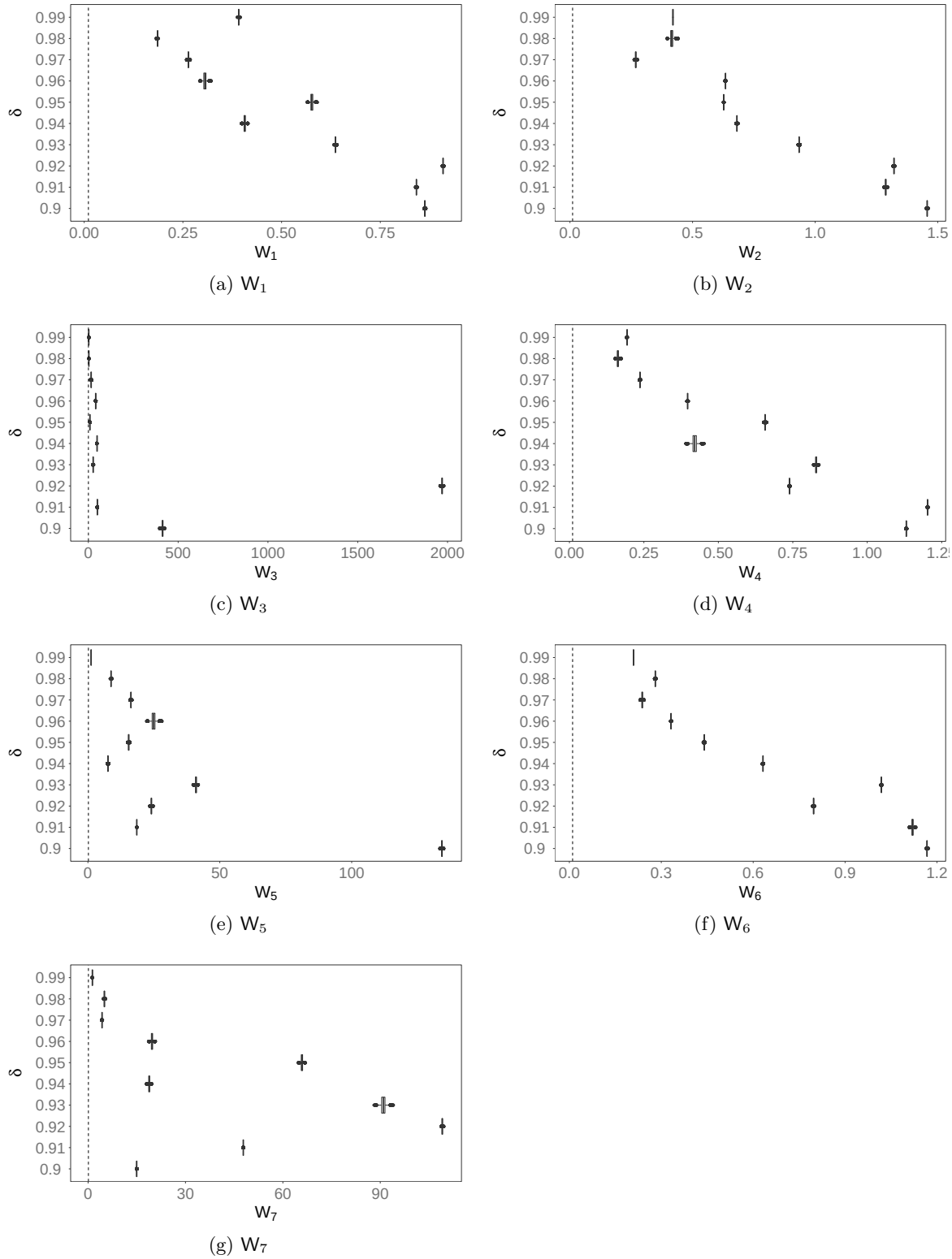
E.2 WC98 data

δ	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$	$\overline{\mathsf{W}}_4$	$\overline{\mathsf{W}}_5$	$\overline{\mathsf{W}}_6$	$\overline{\mathsf{W}}_7$
0.9	0.8628	1.457	413.3846	1.1312	134.1986	1.1675	14.9789
0.91	0.8414	1.2878	50.0254	1.2025	18.5378	1.1202	47.8093
0.92	0.909	1.3216	1969.7943	0.7389	24.0648	0.7973	109.1115
0.93	0.6363	0.9346	26.6734	0.8284	40.9741	1.0184	90.9555
0.94	0.4065	0.6811	48.613	0.4208	7.6053	0.6321	18.8307
0.95	0.5766	0.6269	8.5366	0.657	15.4659	0.4405	65.8818
0.96	0.3057	0.6344	40.8584	0.3968	24.9072	0.3318	19.6853
0.97	0.2639	0.2683	15.7215	0.237	16.3199	0.2379	4.2963
0.98	0.1859	0.4152	2.8881	0.1626	8.7728	0.2805	5.0199
0.99	0.3914	0.420	2.4695	0.1936	1.2208	0.2093	1.3172
PMMH	0.0107	0.0107	0.3288	0.0106	0.1911	0.0106	0.1005

Table E.3: Summary of parameter posterior mean estimation with L&W for the WC98 dataset with varying δ.

δ	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	$\boldsymbol{\theta}_4$	$\boldsymbol{\theta}_5$	$\boldsymbol{\theta}_6$	$\boldsymbol{\theta}_7$
0.9	2136.9928	10620.5383	15579.3281	8520.7728	5633.6589	2109.2161	2179.786
0.91	121.0931	3564.2157	4762.5546	4005.8354	5391.0083	5953.1436	2657.1972
0.92	2141.6233	15111.247	59155.1836	5937.1869	5583.1192	5693.4518	4490.0551
0.93	511.4903	2318.2187	2334.421	2960.5844	3191.5043	2420.1953	1557.9412
0.94	56.595	2549.2358	4609.9489	1315.8808	1043.2483	444.3256	217.3488
0.95	196.7905	730.169	551.4553	1152.1511	1368.1654	1414.8072	587.5594
0.96	964.2013	2017.9215	2207.9498	1634.6758	1551.7697	898.6518	719.8839
0.97	39.4096	849.9581	1361.3209	757.7612	365.2724	148.63	109.2683
0.98	48.5102	418.4741	401.3284	443.4762	328.6819	142.3925	88.0279
0.99	142.0518	153.5734	155.3539	93.8102	127.9376	67.3565	45.5874

Table E.4: Summary of the state posterior mean MSE compared to PMMH using L&W for the WC98 dataset.



(a) W_1 (b) W_2

(c) W_3 (d) W_4

(e) W_5 (f) W_6

(g) W_7

Figure E.2: Parameter posterior estimation using L&W with different δ for the WC98. Vertical dashed line represents PMMH estimated posterior mean

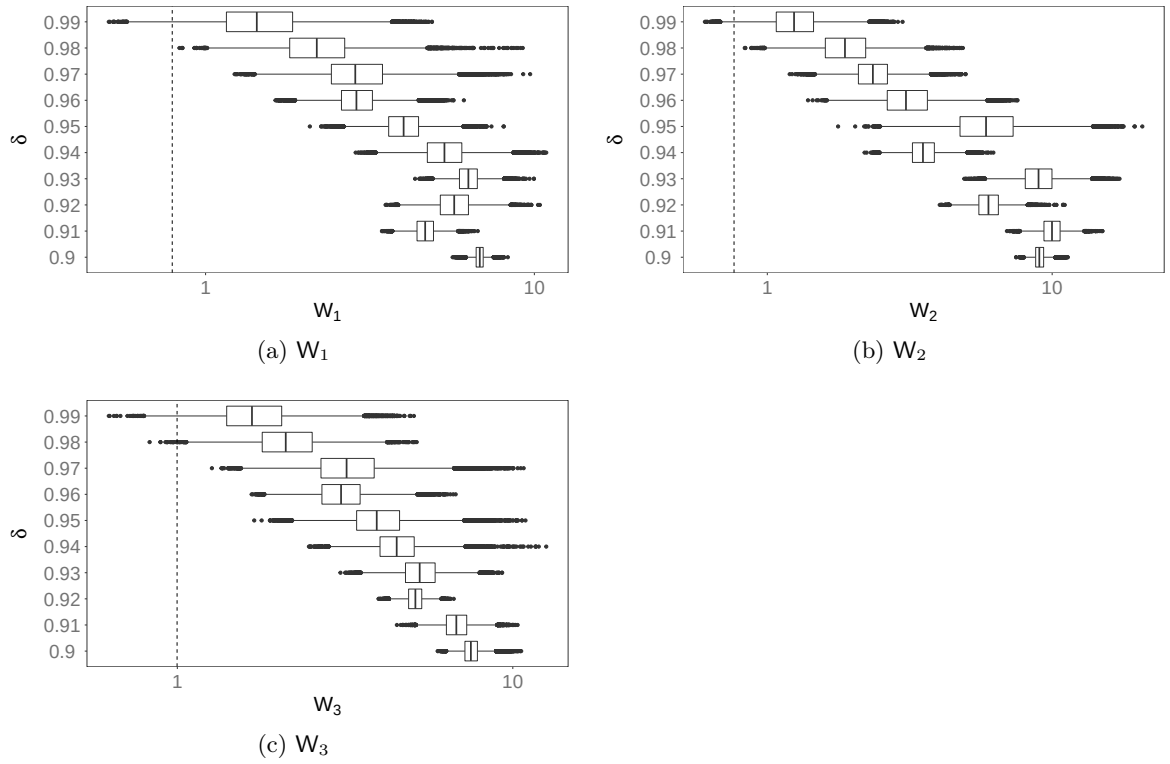
E.3 Airport data

δ	$\overline{\mathsf{W}}_1$	$\overline{\mathsf{W}}_2$	$\overline{\mathsf{W}}_3$
0.9	6.8178	9.0711	7.5618
0.91	4.6793	10.0347	6.8431
0.92	5.7704	6.0262	5.1288
0.93	6.3369	9.0986	5.3638
0.94	5.4373	3.5642	4.6058
0.95	4.0462	6.2241	4.0969
0.96	2.9419	3.2356	3.1434
0.97	3.0322	2.3856	3.3751
0.98	2.306	1.9428	2.1917
0.99	1.5496	1.2956	1.7875
PMMH	0.7917	0.7642	0.9983

Table E.5: Summary of parameter posterior mean estimation with L&W for the airport dataset with varying δ.

δ	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$
0.9	652.1776	460.7062	539.8721
0.91	326.991	282.4767	242.9469
0.92	345.8762	136.1449	202.4889
0.93	203.7074	156.6679	139.4272
0.94	168.3289	169.8873	119.8247
0.95	145.423	171.6073	146.924
0.96	134.0075	78.7561	94.2771
0.97	157.6549	100.2465	102.0575
0.98	79.6999	52.6786	65.0901
0.99	57.6073	41.1823	48.8723

Table E.6: Summary of the state posterior mean MSE compared to PMMH using L&W for the airport dataset.



(a) W_1 (b) W_2

(c) W_3

Figure E.3: Parameter posterior estimation using L&W with different δ for the airport dataset. Vertical dashed line represents PMMH estimated posterior mean

Appendix F

Theoretical results

F.1 Basic probability rules

$$p(A,B) = p(A|B)\,p(B) \tag{F.1}$$

If follows that:

$$\begin{aligned} p(A,B,C) &= p(\{A,B\},C) = p(\{A,B\}|C)\,p(C) && \text{(F.2)} \\ p(A,B,C) &= p(A,\{B,C\}) = p(A|\{B,C\})\,\underbrace{p(B,C)}_{p(B|C)p(C)} \\ &= p(A|B,C)\,p(B|C)\,p(C) && \text{(F.3)} \end{aligned}$$

Equating (F.2) with (F.3), we get:

$$\begin{aligned} p(A,B|C)\,p(C) &= p(A|B,C)\,p(B|C)\,p(C) \Rightarrow \\ p(A,B|C) &= p(A|B,C)\,p(B|C) && \text{(F.4)} \end{aligned}$$

F.1.1 Variance of sum

$$\text{Var}[A+B] = \text{Var}[A] + \text{Var}[B] + 2\text{Cov}[A,B]$$

F.2 Bayes theorem

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)},$$

with $p(B) > 0$.

F.3 Chapman-Kolmogorov

Having a stochastic process as

$$X = \{x_1, x_2, \ldots, x_n\},$$

if we consider the *joint probability* of X as

$$p_{i_1,i_2,\ldots,i_n}(x_1, x_2, \ldots, x_n),$$

the Chapman-Kolmogorov equation is

$$p_{i_1,i_2,\ldots,i_{n-1}}(x_1, x_2, \ldots, x_{n-1}) = \int_{-\infty}^{\infty} p_{i_1,i_2,\ldots,i_n}(x_1, x_2, \ldots, x_n)\, dx_n$$

F.4 Matrix Algebra

F.4.1 Properties of Transpose Matrices

$$\begin{aligned} (A+B)^T &= A^T + B^T \\ (AB)^T &= B^T A^T \end{aligned}$$

F.5 Digamma approximation

If we consider the *digamma function* where

$$\begin{aligned} \gamma(\alpha_t) &= \frac{d \log \Gamma(\alpha_t)}{d\alpha_t} \\ &= \frac{\Gamma'(\alpha_t)}{\Gamma(\alpha_t)} \end{aligned}$$

the derivative, $\gamma'(x)$ as the *trigamma* function, these function can be approximated numerically according to:

$$\begin{aligned} \gamma(x) &\approx \log(x) - \frac{1}{2x} - \frac{1}{12x^2} + \frac{1}{120x^4} - \frac{1}{252x^6} + \ldots \\ &\approx \log(x) - \frac{1}{2x} \\ &\approx \log(x) \end{aligned}$$

and

$$\begin{aligned}\gamma'(x) &\approx \frac{1}{x}+\frac{1}{2x^2}-\frac{1}{6x^3}-\frac{1}{30x^5}+\frac{1}{42x^7}-\frac{1}{30x^9}+\dots \\ &\approx \frac{1}{x}\left(1+\frac{1}{2x}\right) \\ &\approx \frac{1}{x}\end{aligned}$$

Nomenclature

List of Abbreviations

ACF Auto-correlation function

BDLM Binomial Dynamic Linear Model

DGLM Dynamic Generalised Linear Model

EKF Extended Kalman Filter

EM Expectation-Maximisation

ESS Effective sample size

FA Fully adapted

KF Kalman Filter

KF-SVD SVD-based Kalman Filter

LW Liu & West filter

MCMAE Monte Carlo Mean Absolute Error

MH Metropolis-Hastings

MLE Maximum Likelihood Estimation

MSE Mean Squared Error

NDLM Normal Dynamic Linear Model

PoDLM Poisson Dynamic Linear Model

RTS Rauch-Tung-Strieble smoother

SD Standard Deviation

SVD Singular-value decomposition

Nomenclature

$\boldsymbol{\theta}_n$ n^{th} component of the state vector (unless explicitly a state vector at time n)

$\boldsymbol{\theta}_t$ State vector at time t

$\boldsymbol{\theta}_{0:t}$ State vector sequence $\{\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_t\}$

$\boldsymbol{\theta}_{n,t}$ The state vector component n at time t

η_t Natural parameter

λ_t Linear predictor

$\lambda_t^{(i)}$ Unnormalised auxiliary importance weight for particle i at time t

$[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$ Unspecified distribution, with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$

$\mathbf{I}_n$ $n \times n$ identity matrix

$\mathbf{J}_n(\lambda)$ $n \times n$ Jordan block with diagonal elements λ

$\mathcal{D}_t$ The observation sequence $\{y_1, \ldots, y_t\}$

$\mathcal{D}_t^k$ The observation sequence $\{y_{t-k}, \ldots, y_t\}$

$\mathcal{F}(p, h)$ DGLM Fourier seasonal component with period p and h harmonics

$\mathcal{K}(\cdot)$ Deterministic state-sufficient-statistics recursion

$\mathcal{P}(n)$ DGLM n^{th} order polynomial component

A Matrix A

A^T Transpose of matrix A

F DGLM observation matrix

G DGLM system matrix

W DGLM state evolution covariance

W_n n^{th} diagonal element of matrix W

C_t Kalman Filter filtering density covariance at time t.

ϕ_t Dispersion parameter

$\mathbf{m}_t$ Kalman Filter filtering density mean at time t.

$\text{diag}\,(a_1, \ldots, a_n)$ A $n \times n$ square matrix with diagonal elements $a_1, \ldots, a_n$

$\text{E}\,[\text{X}]$ Expectation of X

$\text{Var}\,[\text{X}]$ Variance of X

Θ_t The state sequence $\{\theta_0, \ldots, \theta_t\}$

$\tilde{\lambda}_t^{(i)}$ Normalised auxiliary importance weight for particle i at time t

$\tilde{\mathsf{C}}_t$ Smoothing filtering density covariance at time t.

$\tilde{\mathbf{m}}_t$ Smoothing density mean at time t.

$\tilde{w}_t^{(i)}$ Normalised importance weight for particle i at time t

$g\,(\cdot)$ Link function

$h\,(\cdot)$ Response function

V Normal DLM observation variance

$w_t^{(i)}$ Unnormalised importance weight for particle i at time t

y_t Observation at time t