# Modelling from spatiotemporal data: a dynamic systems approach

A thesis submitted to the University of Sheffield for the degree of Doctor of Philosophy

**Andrew Zammit Mangion**

Department of Automatic Control and Systems Engineering

November 2011

## Acknowledgements

I would like to thank my supervisors: Prof. V. Kadirkamanathan for his helpful insights on spatiotemporal modelling and control and Dr G. Sanguinetti for the unwavering commitment, support and technical help, particularly in the work related to point processes and dual variational filtering.

A special thanks goes to Tara Baldacchino who made my stay in Sheffield so enjoyable and for the lengthy discussions on parameter estimation with EM, to Dr S. Anderson for his help in optimisation, to Andrew Hills for being able to solve just about anything computer-related and to Parham Aram for his insight on non-parametric analysis in IDE-based spatiotemporal systems. I would also like to thank all the other friends in Room 316, in particular Eliza Condrea, Veronica Biga and Geoff Holmes, and Ke Yuan from the University of Southampton who I worked a lot with on temporal point processes. In addition, thanks are due to my house mates and friends who will leave with me several fond memories of Sheffield, and also to my friends and colleagues from Malta who have supported me along the way.

I am indebted to my family; mum and dad for their support and for always welcoming me at home as though I have never left, and my sister Marie and her husband Kenneth. Finally, a very special thanks to Anaïd, for inspiring me in her uncompromising dedication and commitment to research methodology and for her constant support throughout the entire period. Thank you!

# Abstract

Several natural phenomena manifest themselves as spatiotemporal evolution processes. The study of these processes, which aims to increase our understanding of the spatiotemporal phenomena for their prediction and control, requires analysis tools to infer models and their parameters from collected data. Whilst several studies exist on how to model from highly complex patterns characteristic of spatiotemporal processes, an approach which may be readily employed in a wide range of scenarios, such as with systems with different forms of observation processes or time-varying systems, is lacking. This work fills this void by providing a systems approach to spatiotemporal modelling which can be used with continuous observations, point process observations, systems exhibiting spatially varying dynamics and time-varying systems.

The developed methodology builds on the stochastic partial differential equation as a suitable class of models for dynamic spatiotemporal modelling which can easily cater for spatially varying dynamics. A dimensionality reduction mechanism employing frequency methods is proposed; this is used to bring the spatiotemporal system, coupled with the observation process, into conventional state-space form. The work also provides a series of joint field-parameter inference methods which can cater for the vast range of problems under study. Variational techniques are found to be particularly amenable to these kinds of problem and hence a novel dual variational filter is developed to cater for time-varying spatiotemporal systems. The filter is seen to compare favourably with other conventional approaches and to work well on real temporal data sets.

The potential of adopting a systems approach to spatiotemporal modelling is shown on the large-scale *Wikileaks* data set, the *Afghan War Diary*, where it is found that reliable predictions are possible even in complex scenarios. The encouraging results are a strong indication that the adopted approach may be used for large-scale spatiotemporal systems across several disciplines and thus provide a mechanism by which stochastic models are made available for spatiotemporal control purposes.

# Contents

# List of Figures

# List of Tables

# List of Acronyms

| | |
|---|---|
| **AOG** | armed opposition groups |
| **AWD** | Afghan War Diary |
| **CIF** | conditional intensity function |
| **CML** | coupled map lattice |
| **DKF** | dual Kalman filtering |
| **DVBF** | dual VB filter |
| **EKF** | extended Kalman filter |
| **EM** | expectation-maximisation |
| **IDE** | integro-difference equation |
| **i.i.d.** | independently and identically distributed |
| **IMM** | interacting multiple model |
| **IVP** | initial value problem |
| **EEG** | electroencephalography |
| **GMRF** | Gaussian Markov random field |
| **GP** | Gaussian process |
| **GPS** | global positioning system |
| **GRBF** | Gaussian radial basis function |
| **IED** | improvised explosive device |
| **INLA** | integrated nested Laplace approximation |
| **KL** | Kullback-Leibler |
| **KS** | Kolmogorov-Smirnov |
| **LDS** | linear dynamical system |
| **LGCP** | log-Gaussian Cox process |
| **MALA** | Metropolis-adjusted Langevin algorithm |
| **MAP** | maximum-a-posteriori |
| **MC** | Monte Carlo |
| **MCMC** | Markov chain Monte Carlo |
| **ML** | maximum likelihood |

| | |
|---|---|
| **MSE** | mean square error |
| **NARX** | nonlinear autoregressive exogeneous |
| **NTS** | nucleus tractus solitarii |
| **ODE** | ordinary differential equation |
| **PACF** | pair autocorrelation function |
| **PCCF** | pair cross-correlation function |
| **PDE** | partial differential equation |
| **PF** | particle filter |
| **PHD** | probability hypothesis density |
| **PSD** | power spectral density |
| **PSTH** | post-stimulus histogram |
| **RBF** | radial basis function |
| **RHS** | right hand side |
| **RTS** | Rauch, Tung and Striebel |
| **SDE** | stochastic differential equation |
| **SEM** | supplemented EM |
| **SGCP** | sigmoidal-Gaussian Cox process |
| **SIDE** | stochastic integro-difference equation |
| **SIS** | sequential importance sampling |
| **SPDE** | stochastic partial differential equation |
| **SMC** | sequential Monte Carlo |
| **SW** | sliding window |
| **UAV** | unmanned aerial vehicle |
| **USA** | United States of America |
| **VB** | variational Bayes |
| **VBEM** | variational Bayes expectation maximisation |

# Chapter 1

# Introduction

Countless studies serving to improve our quality of life and well-being are concerned with phenomena evolving over both space and time. Some of these spatiotemporal studies are essential in overcoming great challenges and crises faced by humanity today, such as sustainable energy and food security. Application areas of current interest include the following:

i) Wind energy: Wind energy is considered as one of the most sustainable renewable energies. Wind farms are however costly and it is of paramount importance to place them in regions where their potential (load factor) is maximised (see [1] for a critical note on wind farms in the United Kingdom). Strategic placing is possible only if the spatiotemporal behaviour of the wind in a particular region is accurately predicted [2].

ii) Neuroscience: The brain is probably one of the most complex spatiotemporal systems under study today. In addition to the spatiotemporal firing patterns in response to different cues [3], considerable interest is now given to the human brain's transcriptome, i.e. where and when certain genes are expressed in neurodevelopment. Understanding the spatiotemporal dynamics of gene expression in the human brain may further our understanding of certain neurological and psychiatric disorders [4].

iii) Epidemiology: Spatiotemporal tools may be used to study in detail and visualise the dispersion of an epidemic. In doing so underlying factors may reveal why a localised outbreak progressed into an epidemic of international scale, thus assisting in fighting the spread of certain diseases [5]. However epidemiology is not only restricted to the retrospective study of outbreaks: The ability to detect the onset of an epidemic (both in space and in time) may save thousands of lives [6].

iv) Social dynamics: The spatiotemporal behaviour of social phenomena is increasingly becoming a topic of interest [7]. One such recent social phenomenon is the *Arab spring*, a revolutionary 'wave' of protests fuelled by social media networks and exhibiting spatial interactions on a global scale. Another example is armed conflict [8]: In this thesis it is shown how the study of spatiotemporal interactions may be used not only to describe but also predict with confidence measures activity in the near future.

Advances in technology have also led to initiatives to *control* these space-time phenomena in some intelligent way. Control is evident, for example, in the slowing down of a rapidly spreading oil field by using skimmers. In the case of an epidemic, as described above, initiatives may be taken to cease the spread of the disease or at least dampen it enough in a strategic way so as to bring it under control. Another application area where control may prove vital today is in greenhouse gas emissions. The effect of contingency schemes to control the spatiotemporal distribution of carbon dioxide on a global scale, possibly with the use of carbon sinks (through, for example, afforestation) [9], can have enormous implications on the formulation of government policies and economic strategy.

The study of all these phenomena is a highly complex task and usually the analyst is faced with compounding problems. First, the advent of remote sensing technology and advances in telecommunications and data collection methods have resulted in what is commonly termed the *data deluge* [10], an avalanche of information hard to handle due to its sheer size. Second, there is no simple way of *learning* from the data which, without any interpretation, is essentially a collection of thousands of incomprehensible numbers. To get round these two problems there has to be a means with which to summarise the data and with which to provide a level of understanding through description to the end-user.

There exist several ways to tackle these problems, however the literature on appropriate methodology is somewhat fragmented and lacks a method which can easily be extended to cater for different scenarios. This thesis aims to provide a remedy by presenting a novel approach to the study of data which evolves both over space and over time. The proposed methodology is applicable in a variety of scenarios and delivers an approach to learn models which are tailored for prediction and control purposes both from pre-collected data sets and from continuous data streams.

A detailed problem definition and clear research objectives are given in Section 1.3. However, in order to put the motivation in context, this chapter first elaborates on the concept of a model and its estimation in Section 1.1, the fundamental building blocks for any data analysis procedure. Section 1.2 then briefly mentions some model classes commonly associated with spatiotemporal systems and the different types of observations which may be represented by the data. The chapter concludes with a thesis outline.

## 1.1 Introduction to modelling and estimation

### 1.1.1 Deterministic and stochastic models

One of the fundamental building blocks for data analysis is the *model*. A model is a mathematical description which both summarises and describes the behaviour of an observable phenomenon with respect to some independent variables, for example space or/and time. In a few cases the model is *deterministic*, free of unexplained events and perfectly predictable. An example is a model which describes the trajectory of a projectile. In this case $g$, the acceleration due to gravity, is a *parameter* of the model and if it is known, the trajectory of the projectile can be predicted exactly (in the absence of secondary effects). If hypothetically $g$ is unknown, it can be found by studying the trajectory of the projectile in space and time (this is known as the *inverse problem*). Usually, exact measurements in practice are impossible, and what one would typically do in this case is carry out several experiments and find a *likely* value for $g$ by, for instance, taking the mean from the estimates as obtained from each experiment.

Employing statistics to estimate a parameter in this case does not mean that the system itself is *random* in any sense, but rather that the measurements are noisy, and that any (non-systematic) error introduced needs to be filtered out. A separate class of models exists, referred to as *stochastic* models, which intrinsically describe elements which are seemingly random. Stochasticity aims to introduce the concept of uncertainty in the model without obscuring the salient trends, or regularities, of the underlying process. Take for instance the pattern of falling raindrops on a surface; to predict exactly where and when each drop will fall would require an inconceivable amount of deterministic meteorological models, incorporating air pressure, wind speed, water droplet formation and so on. Modelling on these lines is not only infeasible (and virtually impossible) but also unnecessary for many purposes. By studying the pattern of raindrops on a surface one may realise, for instance, that the drops are entirely randomly distributed in space and time (in what we call a Poisson process) and are incident with a very reproducible intensity (e.g. 1 raindrop per cm$^2$ per second). The intensity in itself is highly revealing and may be used for prediction purposes, even though the exact position and time of each incident raindrop will remain uncertain.

As a result of their versatility and their remarkable ability to summarise unexplained effects, stochastic models have seen considerable use in a number of different areas subject to high uncertainty. One of these is financial mathematics, a field pioneered by

Louis Bachelier in his PhD thesis aptly titled *The Theory of Speculation*, first published in 1900 [11]. Stochastic models have also been used for description in fields as diverse as pollution spread [12] and groundwater flow in aquifiers [13]. In these cases the apparent randomness arises from a variety of factors such as climate factors (wind, rainfall) and social factors (market volatility, urbanisation) which are impossible to model deterministically. Nonetheless, by catering for the unexplained effects, it is still possible to use stochastic models for the ultimate purposes of modelling, *prediction* and *control.*

Whilst models, being deterministic or stochastic, seek to further our understanding on the behaviour of a physical phenomenon, their power lies in their predictive abilities, in allowing the possibility to say what is going to happen in the next few seconds, days or years, depending on the application. Knowing what will happen will allow one to anticipate certain effects and even take necessary action to counter them. For deterministic models, prediction is perfect; given an initial condition, the future can be determined with certainty. With stochastic models, the future can only be predicted with *probability.* In the raindrops example one cannot say when and where the next drop will fall, but can predict accurately, and with confidence, how many drops will fall in a certain area in a given time interval. Similarly, in meteorology the average wind speed at a pre-specified location cannot be predicted exactly in advance but, given a good stochastic model, can be predicted with confidence intervals (the average wind speed will be 30kph ±5kph). In essence, stochasticity is a measure of uncertainty; and the more the unexplained effects in the model, the more uncertain the predictions. As a result of their flexibility in systems of high uncertainty, stochastic models will be extensively studied in this work.

### 1.1.2 Parameter estimation

Exploiting the descriptive or predictive powers of both deterministic and stochastic models requires knowledge of the parameters with which they are constructed. In some cases the parameters may be found analytically; for instance one may find the diffusion constant in the heat equation from a material's specific heat capacity, mass density and thermal conductivity. However, this is rarely the case and frequently the model parameters $\boldsymbol{\theta}$ need to be inferred from some set of observations, $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_K\}$. The procedure is known as *parameter estimation.*

There are two widely adopted approaches for parameter estimation. The first attempts to find the *most likely* parameter value from the observations. Formally, it finds the parameter which maximises the *likelihood* $p(\mathcal{Y}|\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* = \arg\max p(\mathcal{Y}|\boldsymbol{\theta}). \tag{1.1}$$

The procedure is also referred to as the *maximum likelihood (ML)* approach. The second approach is known as the *Bayesian* approach and, in contrast, assumes that the parameter is also a random quantity which *a priori* (i.e. without any observations) is distributed according to some prior distribution $p(\boldsymbol{\theta})$.[1] The natural question to ask then is, what is the *posterior* distribution of the parameter given the data set, i.e., what is $p(\boldsymbol{\theta}|\mathcal{Y})$? The answer to this question lies in the use of Bayes' formula which states that

$$p(\boldsymbol{\theta}|\mathcal{Y}) = \frac{p(\mathcal{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{Y})} \propto p(\mathcal{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{1.2}$$

The Bayesian approach thus combines information from the data (likelihood function) and prior information (prior distribution) to provide a distribution over the parameters instead of a point estimate. The distribution of the parameters may be used to provide a confidence of the estimate, which may be useful in many ways.

In practice, the parameters which need to be estimated may require a set of variables $\mathcal{X}$ which are not directly observed (this will be made clear in Section 2.3). In this case the likelihood is computed by *marginalisation* (or the *integrating out*) of $\mathcal{X}$ i.e.

$$p(\mathcal{Y}|\boldsymbol{\theta}) = \int p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta})\mathrm{d}\mathcal{X} = \int p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})p(\mathcal{X}|\boldsymbol{\theta})\mathrm{d}\mathcal{X}. \tag{1.3}$$

In a ML approach, maximisation of (1.3) is typically a much harder problem than that of (1.1). However this problem may be dealt with by using specialised algorithms such as the expectation-maximisation (EM) algorithm [15]. In the Bayesian setting, the posterior over the unknown parameters is given as

$$p(\boldsymbol{\theta}|\mathcal{Y}) = \frac{\int p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})p(\mathcal{X}|\boldsymbol{\theta})\mathrm{d}\mathcal{X} p(\boldsymbol{\theta})}{p(\mathcal{Y})}. \tag{1.4}$$

Likewise, this quantity is rarely available in closed form and requires the use of approximations for its evaluation, such as Markov chain Monte Carlo (MCMC) methods [16].

[1] In this thesis the notation $p(x)$ is used to refer to the *probability distribution of the random variable* $X$, irrespective of whether $X$ is continuous (where $p(\cdot)$ then represents a probability density) or discrete (where $p(\cdot)$ then represents probability mass) [14, Section 2.2].

## 1.2 Spatiotemporal models

A model, with some set of parameters $\boldsymbol{\theta}$ describing a phenomenon occurring in both space and in time is known as a spatiotemporal model. Although several spatiotemporal models are apparent in the literature, they may be largely grouped into two classes, geostatistical models and dynamic models. The differences and commonalities between these two classes are given in Section 1.2.1. A factor which heavily influences both the choice of model and the way in which it is estimated is the nature of the data; whether it is in the form of continuous readings or isolated events. This distinction is elaborated on in Section 1.2.2.

### 1.2.1 Geostatistical and dynamic spatiotemporal models

Geostatistical approaches to modelling data in space and time typically assume the existence of an underlying spatiotemporal Gaussian random function. In such cases the spatiotemporal field is fully specified by its mean and *covariance function* $\phi(\cdot)$ which quantifies how, and to what extent, different points in space and time vary together. For a field $z(\boldsymbol{s}, t)$ with space-time coordinates $(\boldsymbol{s}, t)$, the covariance function is given by

$$\phi(\boldsymbol{r}', t') = \text{cov}(z(\boldsymbol{s}, t), z(\boldsymbol{s} + \boldsymbol{r}', t + t')), \tag{1.5}$$

where $\boldsymbol{r}'$ and $t'$ are spatial and temporal lags respectively. In most practical approaches the function $\phi(\boldsymbol{r}', t')$ is expected to decrease as the distance between two points in space and time increases. Usually a space-time covariance function is hard to construct and frequently a *separable* covariance function is instead employed [e.g. 17, 18] so that:

$$\phi(\boldsymbol{r}', t') = \phi_1(\boldsymbol{r}')\phi_2(t'). \tag{1.6}$$

Dynamic models, on the other hand, consist of difference or differential equations providing an explicit description of the evolution of a spatial field in time. For instance, in discrete time this may be given by

$$z_k = \mathcal{A} z_{k-1} + w_k, \tag{1.7}$$

where $k$ is a discrete temporal index, $\mathcal{A}$ is some possibly non-linear operator and $w_k$ is some additive spatial random field. Dynamic modelling in this way is particularly applicable to practical problems such as prediction and control where the system response at time $k$ may be expressed directly in terms of its state at time $k-1$. Dynamic models are frequently *mechanistic*, in the sense that parameters inferred may have a direct physical meaning, or, at the least bear direct relation to the spatiotemporal behaviour of the system [19].

In a few cases the two models are interchangeable descriptions of the same process [20]. In [21] Storvik shows how, for instance, a separable covariance function with $\phi_2(t') = \theta^{|t'|}, |\theta| < 1$, always corresponds to a field evolving as in (1.7) with $\mathcal{A}z_k = \theta z_k$. In his study he also finds several advantages of dynamic models over geostatistical models:

- Employing signal processing tools with dynamic models renders the parameter estimation process more computationally efficient,
- Inference mechanisms associated with dynamic models can readily handle missing/incomplete data,
- Employed covariance functions may correspond to models of type (1.7) which exhibit unnatural features such as noise fields with negative spatial correlations [21, Example 3].

Despite the advantages of dynamic models, geostatistical models still dominate the literature on spatiotemporal systems. On the other hand, whilst dynamic models have been employed for continuous observations [22, 23], little to no work has yet been done in the context of observations which are available in the form of isolated events.

### 1.2.2 Continuous and point process observations

#### Continuous observations

In the majority of cases, the parameters governing both geostatistical or dynamic models need to be found from a set of continuous readings collected from a finite number of locations, which are either fixed or mobile within the spatial domain.[2] Characteristic data sets include the concentration of airborne particulate matter in the atmosphere obtained from weather stations [12], or temperature and salinity readings from autonomous underwater vehicles [24]. Formally, each observation reading at time $k$, $y_k$, may be considered as a noisy linear mapping $L(\cdot)$ of the spatiotemporal field $z_k$

$$y_k = L(z_k) + v_k, \tag{1.8}$$

[2] Throughout this thesis, the term *continuous readings/observations* will be used to characterise observations taking values on a continuous space, irrespective of whether they are temporally continuous or temporally discrete. When a distinction needs to be made, the temporal nature of the data will be stated explicitly.

where $v_k$ is additive noise, usually assumed to be Gaussian in nature. Other forms for $L(\cdot)$ are mentioned in Section 3.1.2.

Having continuous readings which are linearly related to the field is advantageous in many ways. First, if the spatiotemporal model is also linear in form, standard techniques may be used to estimate the entire field $z_k$ [e.g. 23]. Second, parameter estimation is also facilitated, and in some cases closed-form expressions may be found for all unknown quantities [25]. Third, since $y_k \in \mathbb{R}$, in low noise conditions the readings are highly representative of the spatiotemporal field at that point in space. This contrasts heavily with point process observations.

### Point process observations

The observation process in practice is not restricted to be observed with Gaussian noise; indeed, it is not even restricted to be continuous. A different class of problems exists where only discrete *events*, *emissions* or *points*, are observed; for instance the arrival of customers in a queue [26, Section 1.1] or animal sightings when studying population spread [27]. In this case the observations are denoted as point process observations.

Signal processing with *temporal* point process observations is common in biomedical applications; for instance neural spikes and heart beats are both discrete events in time, separated by seemingly random intervals [28, 29, 30]. The firing intensity of these processes is usually assumed to be representative of some influence which is not directly observable; for example CA1 neurons in the rat hippocampus fire only when the rat is in a certain area, known as the receptive field [31]. The way the receptive field evolves is unknown and thus is usually modelled as a stochastic process. Such a modelling approach has led several authors to call these systems *doubly stochastic point processes* [29, 32]; stochasticity being exhibited both by the precise where and when of an event, and by the varying intensity.

An example of a *spatiotemporal* point process is the example given in the first section, rain falling on a surface. It is likely that the spatiotemporal behaviour of the rain intensity may not be modelled deterministically so that even in this case the process is doubly stochastic. Other examples occurring in practice include earthquake origin time and epicentre [33, 34], crop growth [35] and disease incidence [36, 37]. Occasionally these processes are also referred to as *spatiotemporal Cox processes* in tribute to a pioneer of the theory of point processes, D. R. Cox [26].

It is good to note that spatiotemporal point processes, today, play a major role in disciplines beyond that of spatiotemporal modelling and estimation. One such area is that of multi-target detection, where the points are the targets and the intensity function, also referred to as the *probability hypothesis density (PHD)*, is used to compute the expected number of targets within a given area. The PHD is typically modelled to evolve according to the targets' dynamics and to account for the possibility of spontaneous target dis/appearances. Target detection is carried out with the use of a PHD filter which allows for false positives and false negatives. For further details the interested reader is referred to [38, 39, 40].

## 1.3 Motivation and thesis outline

Data in the form of continuous readings or isolated events may be used to estimate parameters in both types of spatiotemporal models (geostatistical and dynamic) using several methods and approaches. Nonetheless there are limitations with the current state of the art; this section describes these limitations and then proceeds to discuss the research objectives before concluding with a thesis outline and a summary of the contributions of this work.

### 1.3.1 Problem definition

Research on spatiotemporal modelling enjoys a long history and has resulted in several works which tackle problems specific to the nature of the application. However a framework which may be adapted to different problems is lacking in the literature. Specifically, i) a view of how the model fits into a control system is frequently ignored, ii) modelling strategies are highly tailored to the nature of the observation process and considerations to iii) online estimation and iv) spatially varying dynamics are also frequently not touched upon. These issues and their implications will now be discussed in greater detail.

i) From a control perspective, the problem lies with the use of geostatistical models. A large array of works tackle geostatistical models, the parameters of which are learnt from continuous data [17, 18, 2, 41] and from point process data [35, 42, 43], however these types of models have not been used in a control framework to date, presumably because they lack the causality readily apparent in dynamic models. The availability of a dynamic model is of paramount importance if one wishes to employ standard techniques from the control literature, which is almost entirely based on dynamic models. This in itself motivates more research into the inference of dynamic spatiotemporal models.

ii) Spatiotemporal dynamic models have been frequently employed for both deterministic systems observed in noise [44, 45, 46] and for stochastic systems [47, 13, 48].

However point process observations have to date not been treated in the context of dynamic spatiotemporal systems.[3] It is required to fill this void if control methods are eventually to be employed in a point process context. The challenge here is to provide suitable inference mechanisms for this type of observation process.

iii) It is very common to study spatiotemporal problems offline [22, 49] and the concept of parameter estimation in a spatiotemporal system online is still relatively unexplored. Parameters are likely to change in time and these quantities need to be updated 'on-the-go'. Ignoring parameter change or drift may have negative implications in prediction (and hence control) over large horizons, as clearly seen in [43]. It would be highly desirable to employ methods which allow for time-varying parameters in the learning process.

iv) Although spatially varying dynamics are part and parcel of most spatiotemporal systems (as will be evident in Chapter 6), few works tackle the problem of *spatially heterogeneous dynamics* ([19, 50] are exceptions) which, however, could be catered for if some of the underlying mechanisms are understood. As will be seen in Chapter 2 some dynamic models are better than others in retaining mechanistic descriptions. An additional challenge here is to provide an inference mechanism which can cater for spatially varying parameters in the process model.

### 1.3.2 Research objectives

In the light of the above it is deemed necessary to devise an approach which

- uses a dynamic model class flexible enough to describe spatially heterogeneous systems and also to maintain a continuous-space representation. The latter is a requirement for it being able to also handle observations in the form of events which frequently appear in the form of spatial coordinates (from a GPS or other tracking device), but is also necessary in anticipating the inclusion of mobile agents' behaviour within the models.

- allows inference from both continuous and point process observations. This is desirable not only from a practical point of view, but also from the vision that the framework may be used to learn *concurrently* from continuous observations and point process observations (see Chapter 7).

- is practical, allows for fast and efficient estimation methods to be employed and

[3] Notable exceptions are [35], which finds an equivalent geostatistical model for the dynamic system and [19], which assumes a deterministic spatiotemporal evolution process.

provides uncertainty measures over all estimated quantities. Since spatiotemporal processes are usually large-scale, the ability of the methods to handle large quantities of data with ease is of paramount importance. Most of the work in this thesis concentrates on developing these techniques.

- allows for online inference and change detection. In some cases it is desirable to have the ability to detect abrupt, or gradual, changes in the dynamic behaviour of the spatiotemporal system. It is as yet unclear how this would be carried out, particularly in the context of point process observations.

The ultimate aim of this work is to obtain stochastic models from data sets which are amenable for control purposes (control in itself is only briefly touched on in Chapter 7). Control in spatiotemporal domains has been explored [51, 52, 53, 54], however not in the context of stochastic fields. It is not hard to see that many of the proposed approaches for deterministic systems would fail in the event of random disturbances. Control of stochastic spatiotemporal systems is a relatively unstudied topic but bears enormous implications in many application areas of high uncertainty, such as those described in the preface to this chapter. The availability of stochastic spatiotemporal models (and associated inference methodology) tailored for control purposes is a first step in this new area of study.

### 1.3.3 Methodology and thesis outline

In order to facilitate the exposition of the proposed framework the thesis is organised into 7 chapters. Pivotal to satisfying the research objectives is identifying a model class which is amenable to control scenarios (and hence must be dynamic) and easily caters for spatially heterogeneous dynamics. To this end the first part of Chapter 2 compares the properties of the most popular dynamic models and finds that a relatively unstudied class, known as the stochastic partial differential equation (SPDE) (the stochastic counterpart of the more commonly used partial differential equation (PDE)), exhibits the required characteristics. Choosing the SPDE as the class of models under investigation, Chapter 2 proceeds to derive some theoretical properties of this mathematical construct from the literature which allow for its use in practical setups. In particular it proves that the random field may be adequately represented by a finite dimensional object which, in turn, allows the SPDE to be respresented in state-space form. The chapter concludes with an investigation of estimation methodologies associated with the state-space formulation which will feature recurrently throughout the work.

Since SPDEs have not been thoroughly studied in an identification context before, an estimation strategy for this class of models with continuous observations is given in Chapter 3. By adopting the state-space framework, here it is found that estimation methodologies (both Bayesian and non-Bayesian) may be readily applied. Errors introduced by employing a state-space framework are contained through the use of an appropriate basis function selection method. The chapter concludes by studying the effect of poor data quality on the behaviour of some of the estimators.

In Chapter 4 it is seen that, in line with the research objectives, the framework established in Chapter 3 readily extends to point process observations. The same estimation methods implemented in Chapter 3 are tried and tested on a relatively simple spatiotemporal system on a discrete domain. Variational methods emerge as being marginally advantageous over other methods in this scenario. The chapter proceeds to establish a continuous-domain setup which is noted to exhibit considerable advantages over a discrete topology in terms of computational requirements. It concludes with a case study showing, for the first time, the inference of an SPDE from point process observations.

The work in Chapter 3 and Chapter 4 lays the foundation for Chapter 5 which extends the methodology to online scenarios. In light of previous results only variational techniques are implemented. Since online methods in variational methods are scarce, this chapter takes the opportunity to implement a novel variational filter which is well-suited for the task at hand. It is seen that it outperforms conventional methods and provides good results in spatiotemporal systems from both continuous and point process observations. A case study on the latter case is particularly interesting as it clearly shows the advantages of estimating the dynamics of the underlying process, rather than solely the field intensity.

Chapter 6 demonstrates the performance of the developed methods on a real data set. The popular *Wikileaks* data set contains thousands of data points, is event-based and is clearly dynamic in nature. It is found that the event activity follows a particular stochastic model which, akin to a SPDE, can be modelled as an infinite dimensional equation. As a result the finite dimensional methodology and inference methods all readily apply to this scenario. It is found that the dynamic modelling approach with point process observations is capable of providing excellent predictions, even in the presence of large uncertainties.

Chapter 7 concludes with a resumé of the thesis. It places this work in the big picture and discusses future work. Importantly, it discusses the application of this work to a scenario where mobile agents are required to both sense and control a spatiotemporal system governed by SPDEs from continuous observations. It is seen that the developed framework paves the way for further studies on the combined estimation and control of stochastic spatiotemporal systems.

## 1.4 Summary of contributions

The main contribution of this thesis is a new approach to spatiotemporal modelling and estimation fusing together concepts from the literature on SPDEs, signal processing and machine learning. The developed approach is versatile as it can be used with different observation processes (Chapters 3 and 4), heterogeneous systems (Chapters 3, 5 and 6), time-varying systems (Chapter 5) and in the context of spatiotemporal control (Chapter 7). In addition, the research required for the development of this approach has resulted in the following contributions:

i) The problem of SPDE estimation is solved through the development of several estimation methods (Section 3.2) which are applied for the first time in this setting (Section 3.3).

ii) A dimensionality reduction tool for spatiotemporal modelling and estimation is proposed (Section 3.1.3). An empirical study (Section 3.3.4) shows that the parameter bias arising from this approach is contained.

iii) A thorough simulation study (Section 3.4) is used to show and explain why in poorly observed systems the use of variational methods in providing uncertainty measures is somewhat limited.

iv) A novel variational-Laplace algorithm (Section 4.1.2) is especially constructed for the nonlinear smoothing problem in point process systems. The algorithm may find application in other nonlinear dynamic settings.

v) A new signal processing method is developed which is capable of estimating a field's frequency response from point process observations (Section 4.4.1). The availability of frequency information regarding the point process may prove highly beneficial in the context of system identification and control.

vi) A new dual variational filter which allows parameter uncertainty to be propagated in dynamical systems in a deterministic setting is developed (Section 5.2). The filter is shown to outperform other state-of-the-art deterministic filters.

vii) Studies on a real-world data set describing neural responses to taste stimuli show that the developed dual filter is highly suitable for online filtering of biomedical signals and may find considerable application in this regard (Section 5.4.4).

viii) A study on a real-world data set consisting of logged events in a current conflict scenario (Chapter 6) reveals a radically new approach to the study of conflict dynamics. The study also shows that the proposed approach in this thesis is suitable for large-scale systems exhibiting considerable uncertainties.

As a direct consequence of this work, three papers have been published to date [55, 56, 57]. Two further papers are currently under review.

# Chapter 2

# Spatiotemporal and state-space system modelling

This chapter is divided into three parts which together provide the basis for the development of the identification strategies for spatiotemporal systems discussed in the later chapters. The first part of this chapter reviews different types of dynamic spatiotemporal models. Several have been proposed, but here only the most relevant, the coupled map lattice (CML), the integro-difference equation (IDE) and the SPDE are discussed. Their potentials and limitations are highlighted in comparison to each other. Of note in this section is the SPDE, identified as the class of models which describes fields on a continuous spatial domain, is easily set up for spatially heterogeneous dynamics and enjoys good mechanistic descriptions of the underlying process.

The second part, Section 2.2, provides the preliminaries required to bring the SPDE into state-space form, a construct which allows for easy predictions and control implementation. It reviews some theoretical properties of the SPDE which are essential to this thesis, including linear stochastic evolution equations and spatial random fields. These developments will be used in Chapter 3 in conjunction with dimensionality reduction methods commonly employed for the analysis of PDEs, a brief review of which is also given in this section.

Since in this thesis spatiotemporal systems will be represented in the popular state-space model framework prior to analysis, the third part, Section 2.3, is devoted to a review of common methods associated with state-space models such as algorithms for state inference and joint state-parameter inference. Emphasis is given to the algorithms which are employed and investigated throughout the thesis. This section highlights the main differences between the algorithms and provides a simple example which may be used to help interpret the conclusions obtained from results in later chapters.

## 2.1 Spatiotemporal dynamic models

This section briefly reviews three of the most common dynamic spatiotemporal models apparent in the literature, the CML, the IDE and the SPDE.

### 2.1.1 Coupled map lattices

CMLs constitute what are probably the most intuitive family of model structures. They are closely related to cellular automata, with the reduced constraint that the state-space is not necessarily discrete [58]. CMLs are defined to be in discrete-time, discrete-space and may be viewed as a subset of the more general class of lattice dynamic systems [59].

Let $j = 1 \ldots J$ denote a set of lattice points, with each element identifying a discrete location in space. If the field at lattice point $j$ and discrete-time point $k$ is given by $z_{j,k}$, then the temporal evolution at site $j$ may be described through a simple nonlinear mapping $M_j : \mathbb{R}^J \to \mathbb{R}$, so that $z_{j,k+1} = (M_j \mathbf{z}_k)$ with $\boldsymbol{z}_k = [z_{1,k} \ldots z_{J,k}]$. In some cases spatial heterogeneity is modelled [60] but usually a spatially homogeneous process is assumed, so that the mapping dependence on $j$ can be omitted to give a standard nonlinear evolution equation $z_{j,k+1} = (M\boldsymbol{z}_k)$.

The mapping $M$ is what defines the behaviour of the CML. By far the most widely applied mapping is the nearest neighbour coupling map [61, 62, 63, 64], comprising of a local interaction term $f_l$ and a spatial coupling function $f_c$:

$$\begin{aligned} z_{j,k+1} &= f_l(z_{j,k}) + f_c(z_{j-1,k}, z_{j+1,k}) \\ &= (1-\epsilon) f(z_{j,k}) + \frac{\epsilon}{2}\Bigg( f(z_{j-1,k}) + f(z_{j+1,k}) \Bigg), \end{aligned} \tag{2.1}$$

where $f(\cdot)$ is a pre-defined nonlinear function, such as the logistic map $f : z_{j,k} \to 1 - a z_{j,k}^2$, and $\epsilon \in [0, 1]$. Other mappings consider a neighbourhood which extends beyond the nearest neighbours, resulting in what is referred to as 'intermediate range coupling' or 'global coupling' [65]. The neighbourhood adopted significantly affects the output patterns obtained.

Despite the elegant, concise description, CMLs in this form are able to characterise a number of remarkably complex spatiotemporal patterns and have been used to model complicated phenomena such as the sequence of phases and transitions in boiling [66], cloud dynamics [67] and electroencephalography (EEG) signals [68]. The classical way of deriving a CML is through the natural laws governing the spatiotemporal phenomena [62] and thus the resulting model contains parameters which may be directly related to the physical properties of the system. For the case when the mapping $M$ is unknown, or cannot be obtained from first principles, parameter inference and model structure detection may be used [69, 58].

Most CMLs considered in the literature are deterministic. However a *stochastic CML* in which the lattice points are randomly perturbed has been proposed [48]. The stochastic CML is thus a plausible candidate for the framework considered in this thesis since it is dynamic, is highly representative of the system's underlying processes and may be used to represent systems exhibiting large uncertainties. However, in the light of the objectives in Section 1.3.2 it is unsuitable for our purposes for the following reasons:

- Since CMLs are built bottom-up on a discrete grid, observations are restricted to be taken on a regular lattice. Whilst this restriction is impractical in itself, such a construction will not be amenable to control scenarios where mobile agents are employed, or where observations are in the form of isolated events which may, by definition, occur at any point in space.

- Whilst it is possible to define a heterogeneous CML through a spatially varying mapping, it is unclear how this heterogeneity can be parameterised in a convenient way and moreover what inference mechanism may be used to cater for the heterogeneity in parameter estimation.

These limitations warrant the investigation into two other popular dynamic models apparent in the literature, stochastic IDEs, and the somewhat less popular SPDEs.

### 2.1.2 Integro-difference equation models

The main disadvantage of the CML is that it is constructed on a discrete spatial lattice. This may be remedied by employing a continuous-space representation which has received a renewed interest in recent years: the IDE and its stochastic extension, the stochastic integro-difference equation (SIDE).

The deterministic IDE was first introduced as a tool for modelling in ecology by Kot to describe the spread of invading organisms [70, 71]. Kot formulated the IDE by dividing the standard problem of population modelling into two separable stages. The first is known as the *sedentary* stage and is described through a nonlinear map $f(\cdot)$ which typically represents local growth. The second stage is referred to as the *dispersion* stage, described through an integral operator $\mathcal{A}$ and which physically describes diffusion or advection (migration) effects in population.

Let $\boldsymbol{s} \in \mathcal{O} \subset \mathbb{R}^n$, $t \in \mathcal{T} \subset \mathbb{R}^+$, $z(\boldsymbol{s}, t) : \mathcal{O} \times \mathcal{T} \to \mathbb{R}$ be a spatiotemporal field and $k(\boldsymbol{s}, \boldsymbol{r}) : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$ a connectivity kernel. In the simplest case, the IDE defines the field at a point $(t+1)$ through the application of an integral operator

$$z(\boldsymbol{s}, t+1) = \mathcal{A}z(\boldsymbol{s}, t) = \int_{\mathcal{O}} k(\boldsymbol{s}, \boldsymbol{r}) f(z(\boldsymbol{r}, t)) \mathrm{d}\boldsymbol{r}. \tag{2.2}$$

The IDE was subsequently put into a stochastic framework by Wikle [50] by incorporating additive spatial noise through the use of spatial Gaussian processes (GP).

**Definition 2.1 (Gaussian process [72, Section 2.2])** *A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution. It is fully defined by its mean function $\mu(\boldsymbol{s})$ and its covariance function $\Sigma(\boldsymbol{s}, \boldsymbol{r})$ which for a real function $f(\boldsymbol{s})$ are given as $\mu(\boldsymbol{s}) = \mathbb{E}[f(\boldsymbol{s})]$ and $\Sigma(\boldsymbol{s}, \boldsymbol{r}) = \mathbb{E}[(f(\boldsymbol{s}) - \mu(\boldsymbol{s}))(f(\boldsymbol{r}) - \mu(\boldsymbol{r}))]$. A draw $\epsilon(\boldsymbol{s})$ from the GP is then given as*

$$\epsilon(\boldsymbol{s}) \sim \mathcal{GP}(\mu(\boldsymbol{s}), \Sigma(\boldsymbol{s}, \boldsymbol{r})). \tag{2.3}$$

In the SIDE, at each time step the propagated field is superimposed by draws at every time $t$ from a zero-mean spatial GP, $\epsilon(\boldsymbol{s}, t) \sim \mathcal{GP}(0, \Sigma(\boldsymbol{s}, \boldsymbol{r}))$. In compact form the evolution equation of the SIDE is then given as

$$z(\boldsymbol{s}, t+1) = \mathcal{A}z(\boldsymbol{s}, t) + \epsilon(\boldsymbol{s}, t). \tag{2.4}$$

The randomness, introduced through $\epsilon(\boldsymbol{s}, t)$, models uncertainties and caters for any model mismatch or random forcing functions. Since the set generated over time $\{\epsilon(\boldsymbol{s}, t)\}$ is generally assumed to be independently and identically distributed (i.i.d.), the behaviour of the model is largely determined by the form of $f(\cdot)$ and the mixing kernel. For instance $fz = \gamma z, \gamma \in \mathbb{R}^+$ has been used to control and also to allow for exponential growth [50]. Some works have assumed this linear case with $\gamma = 1$ [23, 73]. In EEG studies $f(\cdot)$ is taken to be a sigmoid function [74]. In ecology the standard logistic or Ricker growth models are frequently used [71], however $f(\cdot)$ may easily be set to be of Beverton-Holt, Gompertz or Malthusian form [75].

Under the assumptions that $f(\cdot)$ is a linear function, the mixing kernel wholly defines the dynamic characteristics of the spatiotemporal field. If $k(\boldsymbol{s}, \boldsymbol{r})$ is solely a function of $\boldsymbol{s} - \boldsymbol{r}$, it is invariant under translation and implies that the dynamics of the field are the same throughout space, describing what is termed a *homogeneous field.* If $k(\boldsymbol{s}, \boldsymbol{r}) = k(||\boldsymbol{s} - \boldsymbol{r}||)$, then the kernel is also invariant under rotation and is termed *isotropic.* The shape of $k(\boldsymbol{s}, \boldsymbol{r})$ highly reflects the physical behaviour of the system. Local inhibitions and oscillatory behaviour may be represented through negative lobes and, for instance, if the centre of mass is not placed on the origin (anisotropic kernel), field advection or flow may be described. The appeal of modelling systems with the IDE is that the kernel gives substantial intuitive insight into the system dynamics and indeed, several identification approaches to date have focused on estimating basis functions which shape $k(\boldsymbol{s}, \boldsymbol{r})$ [23, 56, 74, 76]. In the application domain the SIDE has been successfully used for cloud intensity modelling [50], rainfall forecasting [47] and, more recently, for EEG signal modelling [74]. Theoretical extensions and new estimation tools for the SIDE have also been explored fairly recently by Scerri [76].

The favourable properties of the SIDE render it a plausible candidate for consideration in this thesis; it is dynamic, it is defined over a continuous spatial domain and it allows for uncertainty through the additive GP at each time step. However, a key limitation of the SIDE is its inability to provide an underlying physical description of the evolution process. This is best exemplified by noting that the IDE may be simply the integral solution to a PDE. Take for instance the one-dimensional homogeneous heat equation with $D \in \mathbb{R}^+$

$$\frac{\partial z(s,t)}{\partial t} = \frac{D\partial^2 z(s,t)}{\partial s^2}, \quad z(s,0) = z_0(s). \tag{2.5}$$

The Fourier transform with respect to $s$, $Z(\nu, t)$, is given as $\mathrm{d}Z/\mathrm{d}t = (i\nu)^2 DZ(\nu, t)$ which is an ordinary differential equation (ODE) with solution $Z(\nu, t) = Z_0(\nu)e^{-\nu^2 Dt}$. The integral solution is obtained by computing the inverse Fourier transform [77]. Noting that $\mathcal{F}^{-1}(e^{-\nu^2 Dt}) = e^{\frac{-s^2}{4Dt}}/\sqrt{2Dt}$

$$\begin{aligned} z(s,t) &= \frac{1}{\sqrt{2\pi}}\left(z_0(s) * \frac{1}{\sqrt{2Dt}}e^{\frac{-s^2}{4Dt}}\right) \\ &= \frac{1}{\sqrt{4\pi Dt}}\int_{\mathcal{O}} e^{\frac{-(s-s')^2}{4Dt}} z_0(s')ds', \end{aligned} \tag{2.6}$$

where $*$ is the convolution operator. Thus, at $t = 1$, the IDE with a squared exponential function as the connectivity kernel and $f(\cdot) = 1$, constitutes (up to a constant of proportionality) a solution to the heat equation with initial condition $z_0(s)$.

In (2.6) it can be seen that the constant $D$ is embedded in the IDE kernel and thus the physical interpretation of conduction, or heat transfer, is lost in the process. To recover this term it would be required to compare the IDE with the original PDE. This also bears implications on how convenient it is to allow for spatially heterogeneous dynamics. Whilst when employing a PDE spatially varying dynamics may be quantified in some terms (for example through spatially varying conductivity of the medium), heterogeneity in the IDE is hard to represent as this would necessitate allowing for spatially varying redistribution kernels in an arbitrary fashion [50]. The desire for a more mechanistic description and the need for a principled way of representing spatially varying systems leads to the consideration of the PDE, described in the next section.

### 2.1.3 Partial differential equation models

PDEs are continuous-time continuous-space models which have been used extensively in the literature to describe a wide range of natural phenomena. The most popular PDE is undoubtedly the heat equation used to describe heat transfer in (2.5). This is obtained directly from Fourier's law which states that heat energy transfer across a surface of a material is proportional to the temperature gradient across the surface, the constant being a function of specific material properties such as its thermal conductivity and density [78, Chapter 1]. In the same way, the diffusion equation is a natural extension of the heat equation, and the Navier-Stokes and Burger's equation are obtained from the laws of fluid dynamics, just to mention a few. The number of times PDEs have been used for modelling is impressive, finding application in fields as diverse as wildfire control [79], ecology [80], oceanography [81] and flexible structures [45, Section 1.3].

A PDE is formally defined as any equation which involves an unknown function of two or more independent variables and one or more of its partial derivatives [82, Section 1.1]. In spatiotemporal systems the independent variables are restricted to be space and time respectively. Let $s \in \mathcal{O} \subset \mathbb{R}$, $t \in \mathcal{T} \subset \mathbb{R}^+$ and $z(s,t) : \mathcal{O} \times \mathcal{T} \to \mathbb{R}$ be a single-dimension spatiotemporal field for simplicity. Then the general form of the PDE is given by

$$F\left(s, t, z, \frac{\partial}{\partial s}z, \frac{\partial}{\partial t}z, \frac{\partial^2}{\partial s^2}z, \frac{\partial^2}{\partial t^2}z, \frac{\partial^2}{\partial s \partial t}z, \dots\right) = 0. \tag{2.7}$$

If $F(\cdot)$ is a linear function then the PDE is said to be linear, otherwise it is quasilinear or nonlinear. Moreover if $F(\cdot)$ is independent of $s$ and $t$ the system is said to be space and time invariant. Popular examples of linear equations describing an evolution over space and time are the linear transport equation, the heat equation and the wave equation.

PDEs are commonly defined on some *bounded* domain. In this case the PDE formulation needs to include some prescribed conditions for $z$ which must be satisfied on the domain boundary $\partial\mathcal{O}$. These conditions may be either Dirichlet (first-type) requiring $z$ to take on fixed values on $\partial\mathcal{O}$ or Neumann (second-type) which requires $z$ to have fixed derivatives on $\partial\mathcal{O}$. If together with boundary conditions an initial condition is specified, the problem of finding the field $z$ which satisfies the PDE is termed the initial/boundary-value problem.

SPDEs, rather than PDEs, are required as a form of representation when either the forcing term is stochastic in nature [e.g. 83], the initial or boundary conditions are random [84, Section 1.1] or when there is an incomplete knowledge of the physical system. The ensuing flexibility obtained by coupling stochasticity with the deterministic PDE renders the SPDE the most intricate of the three models considered in this section. Prévôt and Röckner [85, Chapter 1] in their opening motivational paragraph state that

> "All kinds of dynamics with stochastic influence in nature or man-made complex systems can be modelled by such equations."

SPDEs have been used for modelling purposes in several application areas. One example of a system governed by an SPDE can be found in the field of hydrology where the groundwater flow in a phreatic aquifer fed by rainfall is described by a linear time-dependent PDE with a random forcing function [13]. Another system is thin-film flow induced by thermal noise [86] in which the authors add thermal noise to the incompressible Navier-Stokes equations to formulate the free surface problem under the influence of fluctuations. Other application areas include the study of neurophysiology [87], turbulence, through the use of the stochastic Burger's equation [88], signal denoising [89] and geophysics [90]. The scope of applications is by no means limited to these and, in principle, every distributed parameter system can be modelled as an SPDE to cater for loss in predictability [84, Section 1.1].

A typical example of a (linear) SPDE is the one-dimensional diffusion equation with a random forcing signal given by

$$\frac{\partial z(s,t)}{\partial t} = \frac{\partial}{\partial s}\left(D(s)\left(\frac{\partial}{\partial s}z(s,t)\right)\right) + \sigma\dot{W}(s,t), \tag{2.8}$$

where $D(s) > 0$, $\sigma \in \mathbb{R}^+$ and $\dot{W}(s,t)$ is space-time noise (to be made precise in Section 2.2). Note that the spatially heterogeneous dynamics are immediately apparent in the spatially varying parameter $D(s)$ which moreover retains physical meaning (e.g. variable conductivity in a metal bar). This contrasts with the IDE where the heterogeneity is implemented in the redistribution kernel based on observed spatiotemporal behaviour [50]. In Chapter 3, and throughout the thesis, it is seen how spatially varying parameters are easily dealt with with minimal effort on the inference procedure.

The reason why SPDEs accomplish the objectives listed in Section 1.3.2 now become apparent. In summary, the formulation of the CML on discrete-space makes it less suitable for point process observations and for situations with mobile agents in the context of spatiotemporal control. On the other hand the SIDE obscures the physical mechanism and presents considerable challenges in describing heterogeneity. The SPDE overcomes all these problems by maintaining a continuous spatial representation, by retaining a physical description of the underlying phenomena and by describing spatially heterogeneous dynamics as a natural part of its formulation.

Choosing the SPDE as the model class for analysis brings with it also other exciting challenges particularly in the context of parameter estimation. Specifically, the literature in this context is mostly set out for dealing with deterministic PDEs observed in noise [91, 92, 45, Chapter 3] and describe methods which do not readily extend to implementation with stochastic fields. Furthermore, to date there has been little effort in the stochastic case with continuous observations [93] and more noticeably none whatsoever with point process observations. This thesis thus turns its focus to the implementation of a systematic framework for the modelling and identification of SPDEs from both continuous observations in Chapter 3 and point process observations in Chapter 4. The methods are extended to the online scenario in Chapter 5.

In order to keep the exposition simple, in this thesis only SPDEs restricted to be linear in form will be considered; the presented approach can also be extended to nonlinear SPDEs with some modifications (see Section 7.2). To meet the objectives set out in Section 1.3.2, some properties of the mathematical construct, in particular the stochastic term, will be discussed in detail in the next section. The discussion is preceded by a brief mention of some popular finite dimensional reduction methods which may be employed to reduce the SPDE, which is inherently infinite dimensional, into a finite dimensional system. The finite dimensional reduction is essential for placing the SPDE into a conventional state-space form which, in turn, will be used as the foundation framework for analysis throughout the thesis.

## 2.2 Basic theory of linear SPDEs

The previous section established the role of the SPDEs in the analysis of spatiotemporal systems and why they are of particular interest in this study. The aim of the present section is to review some preliminaries required for the theoretical framework established in Chapter 3.

### 2.2.1 Model reduction and temporal discretisation of SPDEs

In order to study SPDEs, which are by nature infinite dimensional, it is highly beneficial (and practical) to reduce them into a form amenable to standard signal processing techniques, which are usually tailored for finite dimensional systems. All dimensionality reduction and temporal discretisation methods prevalent in the SPDE literature are extensions of those used for standard PDEs.

The most popular way to discretise PDEs is by using *finite differences* which approximates the temporal and the spatial derivatives of the PDE by difference quotients. For illustration consider the deterministic heat equation of (2.5). Define the discrete operator

$$\mathcal{A}^N z_{j,k} = \frac{1}{\Delta_s^2}(z_{j-1,k} - 2z_{j,k} + z_{j+1,k}), \tag{2.9}$$

where $\Delta_s$ is a fixed-width interval within the spatial domain. Then a *six-point* finite-difference scheme [94, Section 2.6] may be defined through a user-defined parameter $\gamma$ so that (2.5) is approximated to

$$\frac{z_{j,k+1} - z_{j,k}}{\Delta_t} = \mathcal{A}^N(\gamma z_{j,k+1} + (1-\gamma)z_{j,k}), \quad j \in \mathbb{Z}, k \in \mathbb{Z}^+, \tag{2.10}$$

with initial condition $z_{j,0} = z_0(j\Delta_s)$ and where $\Delta_t$ is a fixed-width interval within the temporal domain. Setting $\gamma = 1$ results in what is termed the Euler implicit scheme, $\gamma = 1/2$ the Crank-Nicholson scheme and $\gamma = 0$ the Euler explicit scheme which, written explicitly, yields

$$z_{j,k+1} = z_{j,k}\left(1 - \frac{2\Delta_t}{\Delta_s^2}\right) + \frac{\Delta_t}{\Delta_s^2}(z_{j-1,k} + z_{j+1,k}). \tag{2.11}$$

Letting $2\Delta_t/\Delta_s^2 = \epsilon$ results in the CML (2.1) with $f(\cdot)$ as the identity map. Finite differences approximations have naturally found considerable interest in the SPDE community [95, 96, 97], however, as with CMLs, this representation leads to field description on a discrete domain which is not satisfactory for our purposes.

A different approach, usually associated with spatial dimensionality reduction, is known as the *method of moments* [98, 99, Section 1.3]. Consider the simple linear equation $\mathcal{A}z = f$ for which it is required to find a solution for $z$. By approximately expanding $z$ as a series of $n$ basis functions $\{\phi_i\}_{i=1}^n$ with weights $w_1 \dots w_n$, one obtains an approximation $z \approx \sum_{i=1}^n w_i\phi_i$. Consequently

$$\sum_{i=1}^n w_i \mathcal{A}\phi_i = f. \tag{2.12}$$

The method of moments proceeds by taking the inner product of (2.12) with respect to a set of $m$ testing functions $\{\chi_i\}_{i=1}^m$ to obtain the set of equations

$$\sum_{i=1}^n w_i \langle \chi_j, \mathcal{A}\phi_i \rangle = \langle \chi_j, f \rangle, \qquad j = 1 \dots m. \tag{2.13}$$

The set of equations may be written in matrix form to give

$$\boldsymbol{A}\boldsymbol{w} = \boldsymbol{f}, \tag{2.14}$$

where

$$\boldsymbol{A} = \begin{bmatrix} \langle \chi_1, \mathcal{A}\phi_1 \rangle & \langle \chi_1, \mathcal{A}\phi_2 \rangle & \dots & \langle \chi_1, \mathcal{A}\phi_n \rangle \\ \langle \chi_2, \mathcal{A}\phi_1 \rangle & \langle \chi_2, \mathcal{A}\phi_2 \rangle & \ddots & \langle \chi_2, \mathcal{A}\phi_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \chi_m, \mathcal{A}\phi_1 \rangle & \langle \chi_m, \mathcal{A}\phi_2 \rangle & \dots & \langle \chi_m, \mathcal{A}\phi_n \rangle \end{bmatrix} = [\langle \chi_i, \mathcal{A}\phi_j \rangle]_{i,j=1}^{m,n}, \tag{2.15}$$

and the vectors $\boldsymbol{w} = [w_1, w_2, \dots, w_n]^T$, $\boldsymbol{f} = [\langle \chi_1, f \rangle, \langle \chi_2, f \rangle, \dots, \langle \chi_m, f \rangle]^T$. If $\boldsymbol{A}$ is square and its inverse exists then the required solution is given by solving for $\boldsymbol{w}$ using standard methods.

In the spatiotemporal case, the method of moments results in a linear set of differential equations which may then be represented as a discrete-time state-space model using standard Euler techniques (see Section 3.1). The approximating functions, or basis functions, have to be chosen to exactly satisfy the boundary conditions and suitably approximate the solution [99, Section 1.3]. The choice of functions is not unique and a poor choice effectively results in a bad approximation of the solution. The popular Galerkin method is a special case of the method of moments and is obtained by letting the set of test functions be identical to the set of basis functions, $\{\chi_i\}_{i=1}^m = \{\phi_i\}_{i=1}^n, m = n$. The Galerkin method has been extensively used for both PDE [100, 45, Section 2.2] and SPDE [98, 101, 102, Section 2.2] approximation, although its use has yet to be exploited in the context of data-driven estimation of SPDEs.

The Galerkin approach (and method of moments approaches in general) has some advantages compared to finite difference approaches. In particular it can be used in spaces with complex geometries more easily and can handle Dirichlet boundary conditions systematically by appropriate choice of basis functions. In light of the objectives laid out for this thesis, the Galerkin method is particularly attractive as it allows for continuous-space representations through the use of smooth functions $\{\phi_i\}$. It may hence be employed in the presence of both continuous and point process observations.

The difficulty with employing dimensionality-reduction methods with SPDEs is that of approximating the noise term in an appropriate way [103]. Hence, the following section provides a brief discussion on the stochastic elements of the SPDE and relevant properties which are needed for the Galerkin approximation.

### 2.2.2 Brownian motion and white noise

An integral part to the discussion of SPDEs is the most fundamental stochastic process, *Brownian motion.* Brownian motion, also known as the Wiener process in recognition of Robert Wiener, is defined as follows:

**Definition 2.2 (Brownian motion [104, Section 3.5])** *A real-valued stochastic process* $\beta(t)t \geq 0$ *is called a Brownian motion if it satisfies the following* $\forall 0 \leq t' < t < \infty$

*(i)* $\beta(0) = 0$ *(almost surely),*

*(ii)* $\beta(t) - \beta(t') \sim \mathcal{N}(0, t - t')$*,*

*(iii)* $\beta(t) - \beta(t')$ *is independent of* $\beta(u), 0 \leq u \leq t'$*.*

Consequently $\mathbb{E}[\beta(t)] = 0$ and $\mathbb{E}[\beta(t)^2] = t$. Now, consider the trajectory of a Brownian motion over a small interval $\delta t$ with $\delta\beta = \beta(t + \delta t) - \beta(t)$. Then, from Definition 2.2, $\mathbb{E}[\delta\beta] = 0$ and $\mathrm{var}[\delta\beta] = \delta t$. The quantity $\delta\beta/\delta t$ is thus randomly distributed with mean 0 and variance 1 and is temporally independent. These are the statistical properties of *white noise* $\zeta(t)$ which can therefore be expressed in the limit $\delta t \to 0$ as

$$\zeta(t)\mathrm{d}t = \mathrm{d}\beta(t). \tag{2.16}$$

Note that it would be incorrect to state $\zeta(t) = \mathrm{d}\beta(t)/\mathrm{d}t$ as Brownian motion is not differentiable [104, Section 3.5] although the notation $\zeta(t) \sim \mathrm{d}\beta(t)/\mathrm{d}t$ is sometimes used [104, Section 4.1].

The relationship of Brownian motion with white noise is essential for finding solutions to several stochastic differential equations (SDE). Consider the following simplest of cases expressed in classical form

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = \zeta(t). \tag{2.17}$$

One may exploit the relationship in (2.16) to re-write (2.17) in terms of process increments as

$$\mathrm{d}x(t) = \mathrm{d}\beta(t), \tag{2.18}$$

which obviously means that $x(t)$ follows the trajectory of a (possibly shifted) Brownian motion $\beta(t)$. The importance of this is that by simply changing the representation of (2.17) a solution to the SDE could be found in terms of a well known stochastic process with established statistical properties. The construct (2.18) is well known as a SDE in Itô form. It is the most widely used representation and that which will be used throughout this work. This construct extends to multi-dimensions [105] and to infinite dimensions [106], as discussed in the next section.

### 2.2.3 Linear stochastic evolution equations in infinite dimensions

As explained in Section 2.1.3, an SPDE is any PDE which contains an implicit or explicit form of uncertainty. In this work, for simplicity, only additive random disturbances will be considered. One example is the one-dimensional stochastic heat equation

$$\begin{aligned} \frac{\partial z(s,t)}{\partial t} &= \frac{\partial^2}{\partial s^2} z(s,t) + \sigma_w \dot{W}(s,t) \quad s \in \mathcal{O}, t \in (0,T], \\ z(s,0) &= z_0(s) \quad s \in \mathcal{O}, \\ z|_{\partial\mathcal{O}} &= 0, \end{aligned} \tag{2.19}$$

where the noise term is expressed as the formal time derivative of a real-valued space-time GP $\dot{W} = \partial W/\partial t$ (see [107] for treatment of SPDEs under the classic representation).

Here the abstract representation of (2.19) will be considered, also known as the abstract Itô representation, introduced for scalar processes in Section 2.2.2. Essentially this entails considering the temporal evolution of (2.19) as an SDE and noting that at each time point, the field takes values on a (infinite dimensional) *function space* rather than on the real line. These infinite dimensional evolution equations are in the most general form expressed on some complex Banach space. However, for most applications it is sufficient to consider them on some separable Hilbert space $H$ equipped with inner product $\langle \cdot, \cdot \rangle$ and norm $|| \cdot || = \langle \cdot, \cdot \rangle^{1/2}$. Let $z(t) = z(\cdot, t) \in H$ and $W(t) = W(\cdot, t) \in H$ and introduce a linear differential operator $\mathcal{A} : H \to H$. Then (2.19) can be re-written as the infinite dimensional system

$$\begin{aligned} \mathrm{d}z(t) &= \mathcal{A}z(t)\mathrm{d}t + \sigma_w \mathrm{d}W(t), \\ z(0) &= z_0, \end{aligned} \tag{2.20}$$

where throughout, the domain of the operator is the set of functions in $H$ which satisfy the boundary conditions exactly, written as $D(\mathcal{A}) = \{z \in H : z|_{\partial\mathcal{O}} = 0\} \subset H$.

The deterministic equivalent of (2.20) with $\sigma_w = 0$, $\mathrm{d}z(t) = \mathcal{A}z(t)\mathrm{d}t$, can be recognised as the standard Cauchy problem, or initial value problem (IVP) [106, Appendix A]. For a finite dimensional linear system[1] the solution to the IVP is well known to be $\boldsymbol{z}(t, \boldsymbol{z}_0) = \boldsymbol{T}(t)\boldsymbol{z}_0$ where $\boldsymbol{T}(t) = e^{t\boldsymbol{A}}$. The solution to the IVP in infinite dimensions is thus also of the form of $z(t, z_0) = T(t)z_0, t > 0$. Likewise, for a linear system (i.e. with $\mathcal{A}$ linear) it can be shown that $T(0) = I$, $T(t+s) = T(t)T(s)$, and that the mapping $T(\cdot)z, z \in D(\mathcal{A})$ is continuous in $\mathbb{R}^+$ [108, Chapter 1]. In this case $\{T(\cdot)\}$ is known as a strongly continuous ($C_0$) semigroup of linear operators. The operator $\mathcal{A}$ is then said to

[1] Functions and operators represented by vectors and matrices respectively.

be a *generator* of a $C_0$ semigroup of linear operators.

This work will consider the treatment of operators which can be represented as the sum of known (possibly simpler) linear operators $\{\mathcal{A}_i\}$ weighted by elements in a parameter vector $\boldsymbol{\vartheta} \in \mathbb{R}^d$ so that

$$\mathcal{A} = \sum_{i=1}^{d} \mathcal{A}_i \vartheta_i. \tag{2.21}$$

To denote the dependence on $\boldsymbol{\vartheta}$, the operator $\mathcal{A}$ shall sometimes be written as $\mathcal{A}(\boldsymbol{\vartheta})$. The use of a parameter vector $\boldsymbol{\vartheta}$ is seen to be very useful in describing spatially varying parameters, or spatial heterogeneity, and its estimation would contribute to the identification of the SPDE operator. As a result of linearity the decomposition does not change the properties of the generator. In particular if each $\mathcal{A}_i$ is linear then $\mathcal{A}$ is also linear.

The stochastic element in (2.20) is a generalised form of Brownian motion of Definition 2.2. Before its formal definition, the notion of the trace of an operator is first introduced. Let $\{e_l\}$ be a set of orthonormal basis in $H$. Then the *trace* of a self-adjoint operator $Q$ with eigenvalues $\{\lambda_l\}$ is defined as

$$Tr(Q) = \sum_{l=1}^{\infty} \langle Qe_l, e_l \rangle = \sum_{l=1}^{\infty} \lambda_l, \tag{2.22}$$

where the first equality is a result of Lidskii's theorem [109, pg. 32]. If this quantity exists, then the operator is said to be of *trace class.*

**Definition 2.3** *$W(t)$ is termed a $Q$-Wiener process $W(t), t \geq 0$, if*

*(i)* $W(0) = 0$ almost surely,

*(ii) $W(t)$ has continuous trajectories,*

*(iii) $W(t)$ has independent increments,*

*(iv) $W(t)$ is equipped with some self-adjoint covariance operator $Q$ of trace class where $Q \in L(H)$, where $L(H)$ is the set of linear mappings on $H$.*

**Lemma 2.1 (Expansion of a $Q$-Wiener process [106, Chapter 4] )** *For the $Q$-Wiener process in Definition 2.3, $\mathbb{E}[(W(t) - W(t'))^2] = \mathcal{N}(0, (t-t')Q)$, $t \geq t' \geq 0$ where $\mathcal{N}$ denotes the normal distribution. Moreover, for strictly positive $Q$, $Qe_k = \lambda_k e_k, \lambda_k > 0$ and it can be shown that for arbitrary $t$, $W$ has the expansion*

$$W(t) = \sum_{j=1}^{\infty} \sqrt{\lambda_j} \beta_j(t) e_j, \tag{2.23}$$

*where $\beta_j = \frac{1}{\sqrt{\lambda_j}} \langle W(t), e_j \rangle$ constitutes a set of mutually independent real-valued Brownian* *motions.*

Lemma 2.1 shows that a $Q$-Wiener process can be represented as an infinite sum on $H$. One can already note how an approximate version of the Wiener process can be obtained by truncating the sum in (2.23) from, say, the $(n+1)^{th}$ term on. It is now desirable to analyse the properties of the stochastic process in terms of the expansion; these are given in the following theorem.

**Theorem 2.1 (Properties of $\langle W(t), h\rangle$)** *Consider the stochastic process $\langle W(t), h\rangle$. It can be shown that*

$$\langle W(t), h\rangle = \sum_{j=1}^{\infty} \sqrt{\lambda}_j \beta_j(t) \langle e_j, h\rangle, \qquad h \in H, \tag{2.24}$$

*is a real-valued Wiener process convergent on $H$. Moreover*

$$\mathbb{E}[\langle W(t), h\rangle \langle W(t), g\rangle] = t\langle Qh, g\rangle, \qquad g, h \in H. \tag{2.25}$$

*Proof.* From Lemma 2.1 let $W(t)$ be a $Q$-Wiener represented by (2.23). Then (2.24) is in $\mathbb{R}$ and

$$\mathbb{E}[\mid \langle W(t), h\rangle \mid^2] = \mathbb{E}\left[\sum_{l=1}^{\infty}\sum_{m=1}^{\infty} \sqrt{\lambda_l \lambda_m} \langle \beta_l(t) e_l, h\rangle \langle \beta_m(t) e_m, h\rangle\right], \tag{2.26}$$

where $\beta_l(t)$ and $\beta_m(t)$ are independent Brownian motions and hence $\mathbb{E}[\beta_l(t)\beta_m(t)] = t\delta_{l,m}$, where $\delta_{l,m} = 1$ if $l = m$ and is 0 otherwise. Therefore, by (2.22) and the *Cauchy-Schwarz inequality* (see for instance [110, pg. 7]) which implies that $|\langle e_l, h\rangle| \leq ||e_l|| \cdot ||h||$,

$$\begin{aligned}\mathbb{E} \mid \langle W(t), h\rangle \mid^2 &= t \sum_{l=1}^{\infty} \lambda_l \langle e_l, h\rangle^2 \\ &\leq t||h||^2 \sum_{l=1}^{\infty} \lambda_l < \infty.\end{aligned} \tag{2.27}$$

The proof for the second part of the theorem proceeds on the same lines as above to give

$$\begin{aligned}\mathbb{E}[\langle W(t), h\rangle \langle W(t), g\rangle] &= t \sum_{l=1}^{\infty} \lambda_l \langle e_l, h\rangle \langle e_l, g\rangle \\ &= t \sum_{l=1}^{\infty} \langle Q e_l, h\rangle \langle e_l, g\rangle \\ &= t \sum_{l=1}^{\infty} \langle Qh, e_l\rangle \langle g, e_l\rangle,\end{aligned} \tag{2.28}$$

since $Qe_l = \lambda_l e_l$ and $Q$ is self-adjoint. Now, since for any function $f \in H$, $f = \sum_{l=1}^{\infty} \langle f, e_l \rangle e_l$,

$$t\langle Qh, g\rangle = t\Big\langle \sum_{l=1}^{\infty} \langle Qh, e_l\rangle e_l, \sum_{m=1}^{\infty} \langle g, e_m\rangle e_m \Big\rangle, \tag{2.29}$$

which on expanding gives (2.28),[2] thus completing the proof. ■

$Q$-Wiener processes of Definition 2.3 can be used to represent conventional spatially coloured noise (which unlike white noise is realisable), and are hence those of most interest in this work. In practice, a natural choice for $Q$ is an integral operator with a smooth symmetric kernel as in [111, Section 3.3.2] such that

$$Qu(\boldsymbol{s}) = \int_{\mathcal{O}} k_Q(\boldsymbol{s} - \boldsymbol{r}) u(\boldsymbol{r}) \mathrm{d}\boldsymbol{r}. \tag{2.30}$$

Another type of process, of less practical importance but which has been extensively studied in this area, is known as the *cylindrical* Wiener process denoted here as $W_c(t)$. In cylindrical Wiener processes the restriction of $Q$ being of trace class in Definition 2.3 is lifted [106, Chapter 4]. Take, for instance $k_Q(\boldsymbol{s} - \boldsymbol{r}) = \delta(\boldsymbol{s} - \boldsymbol{r})$ then $Q = I$ (space-time white noise) with $\lambda_k = 1, k = 1, 2, \ldots$. From (2.22) $Tr(Q) = \infty$ and from (2.23) $W_c(t)$ is not $H$-valued.

Such processes are in fact unrealisable and hence of little engineering value. However it is interesting to note that a real-valued representation of $W_c(t)$ can always be derived under projection. This representation in turn also lends itself easily to numerical approximation methods. Again, taking $Q = I$ and using the same reasoning as in the proof of Theorem 2.1, the process

$$\langle W_c(t), h\rangle = \sum_{l=1}^{\infty} \beta_l(t) \langle e_l, h\rangle, \qquad h \in H, \tag{2.31}$$

is a real-valued Wiener process since $\mathbb{E} \mid \langle W(t), h\rangle \mid^2 = t\sum_{l=1}^{\infty} \langle e_l, h\rangle^2 = t||h||^2 < \infty$. Moreover $\mathbb{E}[\langle W_c(t), h\rangle \langle W_c(t), g\rangle] = t\langle h, g\rangle \quad g, h \in H$. The corresponding existence proof for any arbitrary cylindrical process is much more involved and the interested reader is referred to [106, Section 4.3].

Theorem 2.1 will be used for the Galerkin reduction of SPDEs in Chapter 3 and, together with equations defining the observation process, will be used to put the SPDE into state-space form, a representation which lends itself easily to estimation and control. Hence, the next section is devoted to a description of state-space models, and a review

[2] If this is not evident, carry out the expansion with $l, m = 2$ and recall that $\langle e_l, e_m\rangle = \delta_{l,m}$.

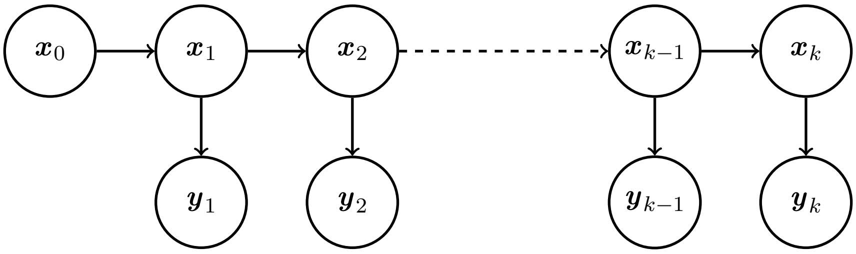


Figure 2.1: Graphical representation of a state-space model showing the evolution of the latent states $\boldsymbol{x}_k$ and the observations $\boldsymbol{y}_k$.

of estimation methods commonly associated with this class of models which will be employed in later chapters.

## 2.3 State-space models

A system is a physical entity whose behaviour can be described by a set of (dynamic) equations which evolve in time. At each instant, a system is said to be in a given *state* and the temporal evolution of the states describe the dynamics of the system. A spatiotemporal system is a system where the state describes spatial characteristics. The coupling of a spatiotemporal system with a stochastic observation process yields the popular stochastic state-space model framework which, to facilitate algorithm development, is typically considered in the discrete-time finite dimensional form [23, 50, 93, 112, 113, 114].

A discrete-time finite dimensional state-space model consists of a real-valued state vector $\boldsymbol{x}_k \in \mathbb{R}^n$ following a first-order Markov process. The sequence of states is not observed directly, but rather, through observations $\boldsymbol{y}_k \in \mathbb{R}^m$, as depicted in Figure 2.3. From the figure several conditional dependencies which facilitate algorithm derivation may be highlighted, for instance that $\boldsymbol{y}_k|\boldsymbol{x}_k$ is conditionally independent of $\boldsymbol{x}_0 \dots \boldsymbol{x}_{k-1}, \boldsymbol{y}_1 \dots \boldsymbol{y}_{k-1}$ or that $\boldsymbol{x}_k|\boldsymbol{x}_{k-1}$ is conditionally independent of $\boldsymbol{x}_0 \dots \boldsymbol{x}_{k-2}, \boldsymbol{y}_1 \dots \boldsymbol{y}_{k-1}$. There is no restraint on the Markovian dynamics and the observation process which may be linear or nonlinear, Gaussian or non-Gaussian. The simplest (and most widely used) system is, however, the linear Gaussian system (or linear dynamical system (LDS)) given as

$$\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{x}_{k-1} + \boldsymbol{w}_k, \qquad \boldsymbol{w}_k \sim \mathcal{N}_{\boldsymbol{w}_k}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \tag{2.32}$$

$$\boldsymbol{y}_k = \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{v}_k, \qquad \boldsymbol{v}_k \sim \mathcal{N}_{\boldsymbol{v}_k}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{v}}), \tag{2.33}$$

where $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is the state transition matrix, $\boldsymbol{C} \in \mathbb{R}^{m \times n}$ is the observation matrix and $\mathcal{N}_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{x} \in \mathbb{R}^n$ denotes the normal distribution of $\boldsymbol{x}$ with mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$

and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$.

One key problem in spatiotemporal systems is the reconstruction of the field in some spatial domain $\mathcal{O}$ at any given time point from some observation process. Obviously if the data is very informative (for instance data obtained with the use of an infrared camera [51]), then the field may be assumed to be known in its entirety without any need for further signal processing. If on the other hand the field is measured at isolated points, as is the case with ocean [24] or neural field [74] sampling, the estimation of the states $\mathcal{X} = \boldsymbol{x}_{0:K} = \{\boldsymbol{x}_0 \dots \boldsymbol{x}_K\}$ (for $K$ subsequent regularly spaced time intervals) from $\mathcal{Y} = \boldsymbol{y}_{1:K} = \{\boldsymbol{y}_1 \dots \boldsymbol{y}_K\}$ for its reconstruction is required. The optimal estimation of $\mathcal{X}$ from some data set is referred to as the *smoothing* problem, explored in Section 2.3.1. If, in addition to $\mathcal{X}$, parameters composing $\boldsymbol{A}$, $\boldsymbol{C}$, $\boldsymbol{\Sigma}_w$, $\boldsymbol{\Sigma}_v$ need to be estimated the problem is referred to as a *joint field-parameter estimation* problem. Several methods for solving this problem exist, including EM, variational Bayes expectation maximisation (VBEM) and MCMC methods, discussed in Sections 2.3.2-2.3.4.

### 2.3.1 Filtering and smoothing

In all problems discussed in this thesis it will be assumed that the spatiotemporal field at the $k^{th}$ time point is reconstructed from $\boldsymbol{x}_k$, which in turn needs to be estimated from a noisy observation process. There are two widely accepted approaches for obtaining the posterior distribution of $\boldsymbol{x}_k$, i.e. the distribution of $\boldsymbol{x}_k$ conditioned on the whole data set $p(\boldsymbol{x}_k|\mathcal{Y})$ [115]. The first is the forward-backward algorithm in which a forward pass (filtering) is followed by a backward pass (smoothing). The second is the two-filter smoother which combines forward messages (identical to those obtained by filtering) with backward messages computed in reverse time to obtain smoothed estimates. Here, and throughout the rest of the thesis, the subscript $k|j$ will be used to denote the estimate at time $k$ from data up to time $j$. Using standard terminology, estimates with subscript $k|k-1$ are termed one-step ahead predictions, $k|k$ filtered estimates and $k|K$ smoothed estimates.

#### Forward-backward smoother

**Forward pass:** Consider the filtered quantity

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k}) = \frac{p(\boldsymbol{y}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1})}{p(\boldsymbol{y}_k|\boldsymbol{y}_{1:k-1})} \propto p(\boldsymbol{y}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1}). \tag{2.34}$$

The term $p(\boldsymbol{y}_k|\boldsymbol{x}_k)$ is the likelihood of $\boldsymbol{x}_k$ and the quantity $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1})$ is the predictive distribution given by

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1}) = \int p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})p(\boldsymbol{x}_{k-1}|\boldsymbol{y}_{1:k-1})\mathrm{d}\boldsymbol{x}_{k-1}. \tag{2.35}$$

If, as in the linear case (2.32) and (2.33), $p(\boldsymbol{y}_k|\boldsymbol{x}_k)$ and $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1})$ are Gaussian densities, then since the product of two Gaussian functions is another Gaussian function [116], the filter estimate can be conveniently written as

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k}) \propto p(\boldsymbol{y}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1}) = \mathcal{N}_{\boldsymbol{x}_k}(\hat{\boldsymbol{x}}_{k|k}, \boldsymbol{\Sigma}_{k|k}). \tag{2.36}$$

Recursive equations for $\hat{\boldsymbol{x}}_{k|k}$ and $\boldsymbol{\Sigma}_{k|k}$ may be found in several engineering textbooks [e.g. 117]. The set of recursions are the governing equations of the popular Kalman filter [118].

If the state evolution equation or the observation equation (Equations (2.32) and (2.33) for the LDS) is non-Gaussian or nonlinear, as is the case with point processes, methods for approximation are required. Several approximate filters exist including those based on linearisation [119], the unscented transform [120], and sequential Monte Carlo (SMC) methods [see 121, 122, for comprehensive reviews]. In SMC methods the filtered distribution is treated as a sum of $N$ particles

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta_{\boldsymbol{x}_k^{(i)}}(\boldsymbol{x}_k), \tag{2.37}$$

where $w_k^{(i)}$ is the weight of the $i^{th}$ particle $\boldsymbol{x}_k^{(i)}$ and $\delta_{\boldsymbol{x}}(\cdot)$ denotes the delta Dirac mass centred at $\boldsymbol{x}$. Using sequential importance sampling (SIS) the weights $w_k^{(i)}$ are computed iteratively as

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\boldsymbol{y}_k|\boldsymbol{x}_k^{(i)})p(\boldsymbol{x}_k^{(i)}|\boldsymbol{x}_{k-1}^{(i)})}{\pi(\boldsymbol{x}_k^{(i)}|\boldsymbol{x}_{0:k-1}^{(i)}, \boldsymbol{y}_{1:k})}, \tag{2.38}$$

where $\pi(\cdot)$ is known as an importance sampling distribution, or a proposal distribution. If further the proposal distribution $\pi(\boldsymbol{x}_k^{(i)}|\boldsymbol{x}_{0:k-1}^{(i)}, \boldsymbol{y}_{1:k}) = p(\boldsymbol{x}_k^{(i)}|\boldsymbol{x}_{k-1}^{(i)})$ is chosen, the weights are updated using solely the likelihood:

$$w_k^{(i)} \propto w_{k-1}^{(i)} p(\boldsymbol{y}_k|\boldsymbol{x}_k^{(i)}). \tag{2.39}$$

In SMC methods the weights $w_k^{(i)}$ quickly become degenerate, requiring what is known as resampling [121]. The resampling procedure, together with the propagation/sampling of the particles using the proposal distribution, renders SMC methods inefficient when compared to deterministic methods, especially in high dimensional spaces.[3]

**Backward pass:** Exploiting the fact that $\boldsymbol{x}_k|\boldsymbol{x}_{k+1}$ is conditionally independent of $\boldsymbol{y}_{k+1:K}$, the backward recursion (or smoothing pass) is given by

$$\begin{aligned} p(\boldsymbol{x}_k|\mathcal{Y}) &= \int p(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}|\mathcal{Y})\mathrm{d}\boldsymbol{x}_{k+1} \\ &= \int p(\boldsymbol{x}_k|\boldsymbol{x}_{k+1}, \mathcal{Y})p(\boldsymbol{x}_{k+1}|\mathcal{Y})\mathrm{d}\boldsymbol{x}_{k+1} \\ &= \int p(\boldsymbol{x}_k|\boldsymbol{x}_{k+1}, \boldsymbol{y}_{1:k})p(\boldsymbol{x}_{k+1}|\mathcal{Y})\mathrm{d}\boldsymbol{x}_{k+1} \\ &= \int \frac{p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k)p(\boldsymbol{x}_{k+1}|\mathcal{Y})}{p(\boldsymbol{x}_{k+1}|\boldsymbol{y}_{1:k})}\mathrm{d}\boldsymbol{x}_{k+1}. \end{aligned} \tag{2.40}$$

Again, in the linear Gaussian case one can write

$$p(\boldsymbol{x}_k|\mathcal{Y}) = \mathcal{N}_{\boldsymbol{x}_k}(\hat{\boldsymbol{x}}_{k|K}, \boldsymbol{\Sigma}_{k|K}). \tag{2.41}$$

Recursive equations for $\hat{\boldsymbol{x}}_{k|K}$ and $\boldsymbol{\Sigma}_{k|K}$ may be found in several texts [e.g. 25]. The set of recursions are the governing equations of the popular Rauch, Tung and Striebel (RTS) smoother [125].

If the state evolution equation or the observation process is non-Gaussian or nonlinear then the RTS smoother ceases to be optimal and the approximative filters listed in the forward pass may be extended into approximative smoothers [e.g. 126, 127, 128].

### Two-filter smoother

The two-filter smoother combines messages obtained from forward recursions and backward recursions to obtain smoothed quantities. This smoother is a result of the factorisation

$$\begin{aligned} p(\boldsymbol{x}_k|\mathcal{Y}) &= p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k}, \boldsymbol{y}_{k+1:K}) \\ &= \frac{p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})p(\boldsymbol{y}_{k+1:K}|\boldsymbol{x}_k, \boldsymbol{y}_{1:k})}{p(\boldsymbol{y}_{k+1:K}|\boldsymbol{y}_{1:k})} \\ &\propto p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})p(\boldsymbol{y}_{k+1:K}|\boldsymbol{x}_k) \\ &= \alpha(\boldsymbol{x}_k)\beta(\boldsymbol{x}_k), \end{aligned} \tag{2.42}$$

[3] The popular belief that particle filters beat the curse of dimensionality is not evidenced in the literature [123, 124].

where $\alpha(\cdot)$ is the forward message and $\beta(\cdot)$ is the backward message.[4] The forward message is identical to the filtered estimate and thus may be computed in the same way as (2.34), re-written for completeness as

$$\begin{aligned}\alpha(\boldsymbol{x}_k) &\propto p(\boldsymbol{y}_k|\boldsymbol{x}_k)\int p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})p(\boldsymbol{x}_{k-1}|\boldsymbol{y}_{1:k-1})\mathrm{d}\boldsymbol{x}_{k-1}\\ &= p(\boldsymbol{y}_k|\boldsymbol{x}_k)\int p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})\alpha(\boldsymbol{x}_{k-1})\mathrm{d}\boldsymbol{x}_{k-1}.\end{aligned} \tag{2.43}$$

The backward message, in turn, is found from

$$\begin{aligned}p(\boldsymbol{y}_{k+1:K}|\boldsymbol{x}_k) &= \int p(\boldsymbol{y}_{k+1:K},\boldsymbol{x}_{k+1}|\boldsymbol{x}_k)\mathrm{d}\boldsymbol{x}_{k+1}\\ &= \int p(\boldsymbol{y}_{k+1:K}|\boldsymbol{x}_{k+1})p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k)\mathrm{d}\boldsymbol{x}_{k+1}\\ &= \int p(\boldsymbol{y}_{k+1}|\boldsymbol{x}_{k+1})p(\boldsymbol{y}_{k+2:K}|\boldsymbol{x}_{k+1})p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k)\mathrm{d}\boldsymbol{x}_{k+1}\\ &= \int p(\boldsymbol{y}_{k+1}|\boldsymbol{x}_{k+1})\beta(\boldsymbol{x}_{k+1})p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k)\mathrm{d}\boldsymbol{x}_{k+1}.\end{aligned} \tag{2.44}$$

Once again, for linear Gaussian systems both $\alpha(\boldsymbol{x}_k)$ and $\beta(\boldsymbol{x}_k)$ are Gaussian so that the product $p(\boldsymbol{x}_k|\mathcal{Y})$ is also Gaussian. For general systems, SMC approaches may be extended to the two-filter case [126]. The two-filter approach allows for parallel implementation but, more importantly, is required for deriving tractable computational updates when computing recursions in a VBEM framework; see [129, Section 5.4.2] for further details.

This sub-section has covered the basics of state estimation in the context of state-space models. If in addition to $\mathcal{X}$ a number of unknown parameters $\boldsymbol{\theta}$ are also required to be estimated, a joint field-parameter estimation algorithm is required. These methods are the focus of the following sub-sections.

### 2.3.2 The EM algorithm

The EM algorithm is a ML estimation algorithm introduced by Dempster, Laird and Rubin [15] for latent (hidden)-data or incomplete-data problems.[5] It was first applied to linear discrete stochastic state-space systems by Shumway and Stoffer [130] and recently in a more general setting by Gibson and Ninness [25]. A comprehensive review on the algorithm and its extensions is given by McLachlan and Krishnan [131] whilst intuitive

[4] For notational convenience $\beta(\cdot)$ will be used to denote both Brownian motion and the backward message; the object it represents will be obvious from the context.

[5] In state-space terminology the latent data are the states $\mathcal{X}$, the incomplete-data the data set $\mathcal{Y}$ and the complete-data the joint $(\mathcal{X},\mathcal{Y})$.

summaries of the algorithm can be found in several sources [132, 133, Chapter 9]. Briefly, EM serves a dual purpose; that of estimating the hidden states $\mathcal{X}$ and that of estimating the unknown parameters $\boldsymbol{\theta}$ using an iterative algorithm. Recall that in the context of spatiotemporal systems represented as state-space models, the hidden states $\mathcal{X}$ represent the field. The ML estimation concerns some parameters $\boldsymbol{\theta}$ governing the dynamics and statistics of the evolution and observation processes.

Let the observed data be denoted as $\mathcal{Y}$. For a given unknown parameter vector $\boldsymbol{\theta}$, the joint (or complete-data) likelihood $p(\mathcal{X}, \mathcal{Y} \mid \boldsymbol{\theta})$ is given by

$$p(\mathcal{X}, \mathcal{Y} \mid \boldsymbol{\theta}) = p(\mathcal{X} \mid \mathcal{Y}, \boldsymbol{\theta}) p(\mathcal{Y} \mid \boldsymbol{\theta}), \tag{2.45}$$

where $p(\mathcal{Y} \mid \boldsymbol{\theta})$ is the observation (or incomplete-data) likelihood. Applying natural logarithms to both sides

$$\ln p(\mathcal{Y} \mid \boldsymbol{\theta}) = \ln p(\mathcal{X}, \mathcal{Y} \mid \boldsymbol{\theta}) - \ln p(\mathcal{X} \mid \mathcal{Y}, \boldsymbol{\theta}). \tag{2.46}$$

From (2.46) it is clear that the latent variables $\mathcal{X}$ need to be marginalised out using expectations in order to obtain an expression in terms of the available data $\mathcal{Y}$ and $\boldsymbol{\theta}$ which can be maximised thereafter. It is not immediately apparent what is the optimal distribution $\tilde{p}(\mathcal{X})$ under which to take expectations.

Adding the right hand side (RHS) of (2.46) with $\ln \tilde{p}(\mathcal{X}) - \ln \tilde{p}(\mathcal{X})$, where $\tilde{p}(\mathcal{X})$ is some free-form distribution, and considering the equality under expectations with respect to $\tilde{p}(\mathcal{X})$

$$\begin{aligned}\ln p(\mathcal{Y} \mid \boldsymbol{\theta}) &= \mathbb{E}_{\tilde{p}(\mathcal{X})}\left[\ln \frac{p(\mathcal{X}, \mathcal{Y} \mid \boldsymbol{\theta})}{\tilde{p}(\mathcal{X})}\right] + \mathbb{E}_{\tilde{p}(\mathcal{X})}\left[\ln \frac{\tilde{p}(\mathcal{X})}{p(\mathcal{X} \mid \mathcal{Y}, \boldsymbol{\theta})}\right] && (2.47)\\ &= \mathcal{L}(\tilde{p}(\mathcal{X}), \boldsymbol{\theta}) + KL(\tilde{p}(\mathcal{X}) || p(\mathcal{X} \mid \mathcal{Y}, \boldsymbol{\theta})). && (2.48)\end{aligned}$$

where $\mathcal{L}(\cdot,\cdot)$ and $KL(\cdot||\cdot)$ are the first and second terms on the RHS of (2.47) respectively. The form of (2.48) is very indicative. In particular, the second term on the RHS is the *Kullback-Leibler (KL) divergence* [133, Section 1.6.1] between $\tilde{p}(\mathcal{X})$ and $p(\mathcal{X} \mid \mathcal{Y}, \boldsymbol{\theta})$. It can be shown that the KL divergence $KL(\cdot||\cdot) \geq 0$ rendering the first term $\mathcal{L}(\tilde{p}(\mathcal{X}), \boldsymbol{\theta})$ a lower bound on the incomplete-data likelihood so that

$$\ln p(\mathcal{Y}|\boldsymbol{\theta}) \geq \mathcal{L}(\tilde{p}(\mathcal{X}), \boldsymbol{\theta}). \tag{2.49}$$

The negative of the lower bound $\mathcal{L}(\tilde{p}(\mathcal{X}), \boldsymbol{\theta})$ is known as the variational free energy in statistical physics [134]. Both the EM algorithm and the VBEM algorithm of Section

2.3.3 attempt to maximise this lower bound in an iterative manner.

The EM algorithm operates by considering an estimate $\boldsymbol{\theta}^{(i)}$ (where $i$ is the iteration number), fixing $\boldsymbol{\theta}$ to this value, and then finding $\tilde{p}(\mathcal{X})^{(i+1)}$ such that the lower bound $\mathcal{L}(\cdot)$ is maximised. One way of carrying out the maximisation is by taking functional derivatives of $\mathcal{L}(\tilde{p}(\mathcal{X}), \boldsymbol{\theta})$ with respect to $\tilde{p}(\mathcal{X})$. The optimal form of the free distribution is then found to be

$$\tilde{p}(\mathcal{X})^{(i+1)} = p(\mathcal{X} \mid \mathcal{Y}, \boldsymbol{\theta}^{(i)}). \tag{2.50}$$

This step is known as the *E-step* and renders (2.48)

$$\begin{aligned}\ln p(\mathcal{Y} \mid \boldsymbol{\theta}^{(i)}) &= \mathcal{L}(\tilde{p}(\mathcal{X})^{(i+1)}, \boldsymbol{\theta}^{(i)}) + KL(\tilde{p}(\mathcal{X})^{(i+1)} || p(\mathcal{X} \mid \mathcal{Y}, \boldsymbol{\theta}^{(i)})) \\ &= \mathcal{L}(\tilde{p}(\mathcal{X})^{(i+1)}, \boldsymbol{\theta}^{(i)}),\end{aligned} \tag{2.51}$$

since the KL divergence between two identical densities is zero. After the E-step the lower bound is therefore equal to the log likelihood. It can also be shown that even the derivatives of the lower bound and the log likelihood are the same at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(i)}$ [131, Section 3.7].

In the next step, known as the *M-step*, $\tilde{p}(\mathcal{X})^{(i+1)}$ is then fixed so that

$$\begin{aligned}\ln p(\mathcal{Y} \mid \boldsymbol{\theta}) &= \mathcal{L}(\tilde{p}(\mathcal{X})^{(i+1)}, \boldsymbol{\theta}) \\ &= \mathbb{E}_{p(\mathcal{X}|\mathcal{Y}, \boldsymbol{\theta}^{(i)})}\left[\ln p(\mathcal{X}, \mathcal{Y} \mid \boldsymbol{\theta})\right] - \mathbb{E}_{p(\mathcal{X}|\mathcal{Y}, \boldsymbol{\theta}^{(i)})}\left[\ln p(\mathcal{X} \mid \mathcal{Y}, \boldsymbol{\theta}^{(i)})\right],\end{aligned} \tag{2.52}$$

where the second term is the entropy of $p(\mathcal{X} \mid \mathcal{Y}, \boldsymbol{\theta}^{(i)})$, is independent of $\boldsymbol{\theta}$ and can hence be treated as a constant. The maximisation problem is therefore reduced to maximising the first term in the RHS of (2.52) so that

$$\boldsymbol{\theta}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathcal{X}|\mathcal{Y}, \boldsymbol{\theta}^{(i)})}\left[\ln p(\mathcal{X}, \mathcal{Y} \mid \boldsymbol{\theta})\right]. \tag{2.53}$$

The new parameter estimate $\boldsymbol{\theta}^{(i+1)}$ makes the KL divergence in (2.48) non-zero so that the E-step and M-step have to be repeated until a stopping condition is met, say $||\boldsymbol{\theta}^{(i+1)} - \boldsymbol{\theta}^{(i)}|| < \epsilon$, where $|| \cdot ||$ denotes the usual Euclidean norm and $\epsilon$ is a pre-defined threshold. When this condition is satisfied, the EM algorithm is said to have converged.

**Theorem 2.2** *The EM algorithm is guaranteed to converge to a fixed point $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ for which $\ln p(\mathcal{Y} \mid \boldsymbol{\theta}^*)$ is a stationary value [133, Section 9.4].*

*Proof.* Since the distribution $p(\mathcal{Y}|\boldsymbol{\theta})$ is positive, then under the mild assumption of boundedness, i.e. $\sup\{p(\mathcal{Y}|\boldsymbol{\theta}^{(i)})\} < \infty$, it is sufficient to show that

$$p(\mathcal{Y}|\boldsymbol{\theta}^{(i)}) \geq p(\mathcal{Y}|\boldsymbol{\theta}^{(i-1)}), \qquad i > 0, \tag{2.54}$$

as this implies that $\lim_{i\to\infty} p(\mathcal{Y}|\boldsymbol{\theta}^{(i)}) = p(\mathcal{Y}|\boldsymbol{\theta}^{*})$. For $i > 0$, as a result of the M-step,

$$\mathcal{L}(\tilde{p}(\mathcal{X})^{(i)}, \boldsymbol{\theta}^{(i)}) \geq \mathcal{L}(\tilde{p}(\mathcal{X})^{(i)}, \boldsymbol{\theta}^{(i-1)}). \tag{2.55}$$

Moreover, from (2.52) and (2.48)

$$\ln p(\mathcal{Y}|\boldsymbol{\theta}^{(i-1)}) = \mathcal{L}(\tilde{p}(\mathcal{X})^{(i)}, \boldsymbol{\theta}^{(i-1)}) \tag{2.56}$$

$$\begin{aligned}\ln p(\mathcal{Y}|\boldsymbol{\theta}^{(i)}) &= \mathcal{L}(\tilde{p}(\mathcal{X})^{(i)}, \boldsymbol{\theta}^{(i)}) + KL(\tilde{p}(\mathcal{X})^{(i)}||p(\mathcal{X} \mid \mathcal{Y}, \boldsymbol{\theta}^{(i)}))\\ &\geq \mathcal{L}(\tilde{p}(\mathcal{X})^{(i)}, \boldsymbol{\theta}^{(i-1)}) + KL(\tilde{p}(\mathcal{X})^{(i)}||p(\mathcal{X} \mid \mathcal{Y}, \boldsymbol{\theta}^{(i)})).\end{aligned} \tag{2.57}$$

Equation (2.54) thus follows from the positiveness of KL divergence. ■

It should be noted that although $\{p(\mathcal{Y}|\boldsymbol{\theta}^{(i)})\}$ converges monotonically to a stationary point, this point may well be a saddle point. To circumvent this problem certain regularity conditions were imposed by Wu [135, 131, Section 3.4.1] to ensure that $\boldsymbol{\theta}^{*}$ is a maximiser of $p(\mathcal{Y}|\boldsymbol{\theta})$. Since no such problems were encountered in this work, this issue is not discussed further. A summary of this algorithm is given in Algorithm 2.1.

**Algorithm 2.1** The EM algorithm

```
Input: Data set $\mathcal{Y}$, initial parameter estimate $\boldsymbol{\theta}^{(0)}$.
$i = 0$
while (not converged)
    $\tilde{p}(\mathcal{X})^{(i+1)} = p(\mathcal{X} \mid \mathcal{Y}, \boldsymbol{\theta}^{(i)})$                     E-step
    $\boldsymbol{\theta}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}[\ln p(\mathcal{X}, \mathcal{Y} \mid \boldsymbol{\theta})]$     M-step
    $i = i + 1$
Output: $\tilde{p}(\mathcal{X})^{(i)}$, $\boldsymbol{\theta}^{(i)}$.
```

**MAP-EM** Maximum-a-posteriori (MAP)-EM is a semi-Bayesian approach to parameter estimation through the consideration of a parameter prior distribution $p(\boldsymbol{\theta})$. Since

$$p(\boldsymbol{\theta}|\mathcal{Y}) \propto p(\mathcal{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{2.58}$$

the only modification in the algorithm is the inclusion of the prior in the M-step so that

$$\boldsymbol{\theta}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} \left[\ln p(\boldsymbol{\theta}) + \mathbb{E}_{p(\mathcal{X}|\mathcal{Y},\boldsymbol{\theta}^{(i)})} \left[\ln p(\mathcal{X}, \mathcal{Y} \mid \boldsymbol{\theta})\right]\right]. \tag{2.59}$$

The MAP-EM algorithm is considered only semi-Bayesian since at each step the probability mass of the posterior distribution is focused at its mode, and can hence still be treated as a point estimate. This, in many cases, is unrepresentative of the true distribution. The VBEM algorithm, described next, overcomes this problem by maintaining the distributional properties of the parameter estimates throughout the E-step.

### 2.3.3 The VBEM algorithm

VBEM is an elegant framework for analytic computations of approximate posterior distributions over latent variables and parameters [136, 137]. The posterior distributions are computed using iterations (coined Iterative VB in [14, Section 1.2]), in a similar way as the EM algorithm, and its convergence is guaranteed. Whilst inheriting the advantages of being a Bayesian approach, the method is deterministic, i.e. no sampling is required, and for a given set of data, likelihood, and prior distribution, the approximate posterior distribution is unique. Hence the VB method is very fast when compared to MCMC methods such as that described in Section 2.3.4. It has consequently seen wide applicability in a wide range of problems such as the modelling of the cell's regulatory network [138, 139], vision tracking [140], blind source separation [141] and neuroimaging [142].

The VB method hinges on finding a convenient functional form for approximating the joint posterior distribution $p(\mathcal{X}, \boldsymbol{\theta}|\mathcal{Y})$. Usually the approximation is carried out using conditionally independent distributions $\tilde{p}(\mathcal{X})$ and $\tilde{p}(\boldsymbol{\theta})$ so that $p(\mathcal{X}, \boldsymbol{\theta}|\mathcal{Y}) \approx \tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta})$. Throughout this work $\tilde{p}(\mathcal{X})$ and $\tilde{p}(\boldsymbol{\theta})$ will be referred to as the *variational posterior distributions*. The forms of these distributions are obtained by maximising the free energy using functional derivatives, much in the same way as in the E-step of the EM algorithm.

Rather than being based on the log incomplete-data likelihood (2.46), the VB method is set up using the log evidence[6]

$$\ln p(\mathcal{Y}) = \ln p(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta}) - \ln p(\mathcal{X}, \boldsymbol{\theta}|\mathcal{Y}). \tag{2.60}$$

[6] The term *evidence* arises from the use of the marginal likelihood $p(\mathcal{Y})$ as evidence for a particular model in model selection procedures. In such cases $p(\mathcal{Y})$ is usually explicitly written as $p(\mathcal{Y}|\mathcal{M}_i)$ where $\mathcal{M}_i$ is a candidate model.

This can be represented much in a similar way as (2.48)

$$\begin{aligned}\ln p(\mathcal{Y}) &= \mathbb{E}_{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta})}\left[\ln \frac{p(\mathcal{X},\mathcal{Y},\boldsymbol{\theta})}{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta})}\right] + \mathbb{E}_{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta})}\left[\ln \frac{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta})}{p(\mathcal{X},\boldsymbol{\theta}\mid\mathcal{Y})}\right] \\ &= \mathcal{L}(\tilde{p}(\mathcal{X}),\tilde{p}(\boldsymbol{\theta})) + KL(\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta})||p(\mathcal{X},\boldsymbol{\theta}|\mathcal{Y})). \end{aligned} \tag{2.61}$$

As in EM, the first term of (2.61) $\mathcal{L}(\tilde{p}(\mathcal{X}),\tilde{p}(\boldsymbol{\theta}))$ (the negative of the variational free energy), is a lower bound on the log evidence so that

$$\ln p(\mathcal{Y}) \geq \mathcal{L}(\tilde{p}(\mathcal{X}),\tilde{p}(\boldsymbol{\theta})). \tag{2.62}$$

The forms of the variational posterior distributions which maximise the lower bound are given in the following theorem.

**Theorem 2.3 (VB variational posteriors)** *Given a data set $\mathcal{Y}$ and its log evidence of the form (2.60), the maximum of the lower bound $\mathcal{L}(\tilde{p}(\mathcal{X}),\tilde{p}(\boldsymbol{\theta}))$ is reached for*

$$\tilde{p}(\mathcal{X}) \propto \exp\Big(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta})}[\ln p(\mathcal{X},\boldsymbol{\theta},\mathcal{Y})]\Big), \tag{2.63}$$

$$\tilde{p}(\boldsymbol{\theta}) \propto \exp\Big(\mathbb{E}_{\tilde{p}(\mathcal{X})}[\ln p(\mathcal{X},\boldsymbol{\theta},\mathcal{Y})]\Big). \tag{2.64}$$

*Proof.* Taking the functional derivative of

$$\mathcal{L}(\tilde{p}(\mathcal{X}),\tilde{p}(\boldsymbol{\theta})) = \mathbb{E}_{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta})}\left[\ln \frac{p(\mathcal{X},\mathcal{Y},\boldsymbol{\theta})}{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta})}\right], \tag{2.65}$$

with respect to $\tilde{p}(\mathcal{X})$ yields

$$\begin{aligned}\frac{\delta\mathcal{L}(\tilde{p}(\mathcal{X}),\tilde{p}(\boldsymbol{\theta}))}{\delta\tilde{p}(\mathcal{X})} &= \int \tilde{p}(\boldsymbol{\theta})\ln\left[\frac{p(\mathcal{X},\boldsymbol{\theta},\mathcal{Y})}{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta})}\right]\mathrm{d}\boldsymbol{\theta} - 1 \\ &= \int \tilde{p}(\boldsymbol{\theta})\ln p(\mathcal{X},\boldsymbol{\theta},\mathcal{Y})\mathrm{d}\boldsymbol{\theta} - \ln\tilde{p}(\mathcal{X}) - C_0 \\ &= 0, \end{aligned} \tag{2.66}$$

for a maximum. Here $C_0 = 1 - \text{entropy}(\tilde{p}(\boldsymbol{\theta}))$. Rearranging gives (2.63). The proof for (2.64) follows by symmetry. ■

The VBEM algorithm operates by considering a parameter distribution $\tilde{p}(\boldsymbol{\theta})^{(i)}$ and then finding $\tilde{p}(\mathcal{X})^{(i+1)}$ such that the lower bound is maximised. Next, $\tilde{p}(\mathcal{X})^{(i+1)}$ is fixed and $\tilde{p}(\boldsymbol{\theta})^{(i+1)}$ is found such that the lower bound is maximised. The VBEM algorithm converges to a local maximum in $\mathcal{L}(\tilde{p}(\mathcal{X}),\tilde{p}(\boldsymbol{\theta}))$, since, i) as a result of subsequent maximisations,

$$\begin{aligned}\mathcal{L}(\tilde{p}(\mathcal{X})^{(i+1)}, \tilde{p}(\boldsymbol{\theta})^{(i+1)}) &\geq \mathcal{L}(\tilde{p}(\mathcal{X})^{(i+1)}, \tilde{p}(\boldsymbol{\theta})^{(i)}) \\ &\geq \mathcal{L}(\tilde{p}(\mathcal{X})^{(i)}, \tilde{p}(\boldsymbol{\theta})^{(i)}),\end{aligned} \tag{2.67}$$

and ii) $\mathcal{L}(\tilde{p}(\mathcal{X})^{(i+1)}, \tilde{p}(\boldsymbol{\theta})^{(i+1)})$ is bounded from above by $\ln p(\mathcal{Y})$ in (2.61). As in the EM algorithm, convergence may be assessed by monitoring, for instance, the change in the mean of the parameter posterior distribution across consecutive iterations. A summary of the VBEM algorithm is given in Algorithm 2.2.

**Algorithm 2.2** The VBEM algorithm

```
Input: Data set $\mathcal{Y}$, initial parameter variational posterior distribution $\tilde{p}(\boldsymbol{\theta})^{(0)}$.
$i = 0$
while (not converged)
    $\tilde{p}(\mathcal{X})^{(i+1)} \propto \exp\left(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta})^{(i)}}[\ln p(\mathcal{X}, \boldsymbol{\theta}, \mathcal{Y})]\right)$        VBE-step
    $\tilde{p}(\boldsymbol{\theta})^{(i+1)} \propto \exp\left(\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}[\ln p(\mathcal{X}, \boldsymbol{\theta}, \mathcal{Y})]\right)$        VBM-step
    $i = i + 1$
Output: $\tilde{p}(\mathcal{X})^{(i)}$, $\tilde{p}(\boldsymbol{\theta})^{(i)}$.
```

The VBEM exhibits a lot of similarity with the conventional EM algorithm. A significant difference, however, is that $\tilde{p}(\mathcal{X})^{(i+1)}$ is found using the expectations of $\boldsymbol{\theta}$ rather than solely its ML point estimate. The two methods will thus differ considerably when, for instance, the posterior mode differs from the posterior mean. This is an advantage of VBEM which through averaging does not give too much importance to the mode of the parameter posterior distribution. This feature makes it ideal for skewed unimodal distributions such as those generated by point process systems, as discussed in Chapter 4. From a different perspective one may view the EM algorithm as a special case of VBEM known as *functionally constrained* VB [143], where the parameter variational posterior is constrained to a point mass function $\tilde{p}(\boldsymbol{\theta})^{(i)} = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)})$ with $\boldsymbol{\theta}^{(i)}$ at the mode of the posterior distribution [144]. Using terminology from classical control, the EM algorithm capitalises on a certainty equivalence property whilst the VBEM algorithm is more 'cautious' by incorporating knowledge of second order and higher moments in state estimation [145, 146, Section 6.4].

### 2.3.4 Gibbs sampling

MCMC methods are by far the most widely used class of distributional approximation methods. The success of MCMC methods is attributed to the following facts: i) convergence to the target posterior distribution is guaranteed [147, Chapter 6], ii) they are generally simple to implement, iii) they can be applied to most models with ease and iv) recent advances in technology and parallel computing have made them applicable to large-scale inference problems.[7] Briefly, MCMC methods attempt to generate a Markov chain, the stationary distribution of which is the desired posterior distribution (see [16, Section 7.1] for a detailed overview). The two arguably most prevalent sampling approaches are the Metropolis-Hastings algorithm [150] and the Gibbs sampler [151]. The latter is particularly useful when the functional form of the joint posterior distribution $p(\mathcal{X}, \boldsymbol{\theta}|\mathcal{Y})$ is not known, or hard to sample from, but where the conditional densities $p(\mathcal{X}|\boldsymbol{\theta}, \mathcal{Y})$ and $p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y})$ are known, or easy to sample from. Since this is generally the case with state-space models, the Gibbs sampler has found wide applicability in this scenario [152, 153].

Consider a parameter sample $\boldsymbol{\theta}^{(i)}$. A basic two-stage sampler (see Algorithm 2.3) generates a state sample $\mathcal{X}^{(i+1)}$ from $\mathcal{X}_{i+1} \sim p(\mathcal{X}|\boldsymbol{\theta}^{(i)}, \mathcal{Y})$, and then $\boldsymbol{\theta}^{(i+1)}$ from $\boldsymbol{\theta}_{i+1} \sim p(\boldsymbol{\theta}|\mathcal{X}^{(i+1)}, \mathcal{Y})$. The procedure is then repeated for $N$ times or until some criterion is met. All statistical moments of the individual and joint posterior distributions may be evaluated from the stationary distributions of the resulting Markov chains. In particular one has that for $N$ samples

$$\mathbb{E}_{p(\mathcal{X}|\mathcal{Y})}[f(\mathcal{X})] \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathcal{X}^{(i)}), \tag{2.68}$$

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{Y})}[f(\boldsymbol{\theta})] \approx \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}^{(i)}), \tag{2.69}$$

$$\mathbb{E}_{p(\mathcal{X},\boldsymbol{\theta}|\mathcal{Y})}[f(\mathcal{X}, \boldsymbol{\theta})] \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathcal{X}^{(i)}, \boldsymbol{\theta}^{(i)}). \tag{2.70}$$

In practice the first samples need to be excluded from the approximations (2.68)-(2.70) to avoid bias introduced by possibly inappropriate initial conditions. This process is referred to as burn-in.

Despite the theoretical and practical advantages of MCMC techniques, they are stochastic approximation methods; the final distributional approximation is determined by the paths of the Markov chains, which by definition are random. Associated methods are usually computationally inefficient and it is hard to assess when the chain has

[7] This is particularly so with the increased availability and performance of extremely powerful graphical processing units, such as those employing the NVIDIA® CUDA™ parallel computing architecture [148, 149].

converged with acceptable error to the distribution of interest. These limitations, which are somewhat exacerbated in high dimensional systems [154] such as spatiotemporal systems, are what drives research into approximate deterministic inference methods such as EM and VBEM.

**Algorithm 2.3** The two-stage Gibbs sampler

```
**Input:** Data set $\mathcal{Y}$, initial parameter sample $\boldsymbol{\theta}^{(0)}$.
$i = 0$
**while** (not converged)
    Sample $\mathcal{X}^{(i+1)}$ from $p(\mathcal{X}|\boldsymbol{\theta}^{(i)}, \mathcal{Y})$
    Sample $\boldsymbol{\theta}^{(i+1)}$ from $p(\boldsymbol{\theta}|\mathcal{X}^{(i+1)}, \mathcal{Y})$
    $i = i + 1$
**Output:** $\{\mathcal{X}^{(t)}\}_{t=1}^{i}, \{\boldsymbol{\theta}^{(t)}\}_{t=1}^{i}$
```

### 2.3.5 Comparative study: the normal distribution with unknown mean and precision

The operational differences between the above joint field-parameter estimators are easily studied when implementing them on a static model. In this study, $N$ i.i.d. data points $\mathcal{Y}$ were generated from a standard normal distribution $\mathcal{N}(\mu, 1/\tau)$ with unknown mean $\mu$ and unknown precision $\tau$. The prior distributions are given as

$$p(\mu) = \mathcal{N}_\mu(\mu', \sigma_\mu^2), \qquad p(\tau) = \mathcal{G}a_\tau(\alpha_\tau, \beta_\tau), \tag{2.71}$$

where $\mathcal{G}a_\tau(\alpha_\tau, \beta_\tau)$ denotes the Gamma distribution in $\tau$ with shape parameter $\alpha_\tau$ and inverse scale (rate) parameter $\beta_\tau$. The full joint posterior for this problem is

$$p(\mu, \tau|\alpha_\tau, \beta_\tau, \mu', \sigma_\mu^2, \mathcal{Y}) \propto \mathcal{G}a_\tau(\alpha_\tau, \beta_\tau)\mathcal{N}_\mu(\mu', \sigma_\mu^2) \prod_{j=1}^{N} \sqrt{\tau} \exp\left(-\frac{\tau}{2}(y_j - \mu)^2\right). \tag{2.72}$$

Since the normalisation constant in (2.72) cannot be found analytically, the posterior needs to be approximated using, for instance, a Gibbs sampler, the VBEM algorithm or the EM algorithm. These three methods will now be applied to this problem and benchmarked against the posterior evaluated on a very fine grid (this computation is only plausible in simple cases such as that presented here), which will be referred to as the 'true posterior'. For consistency with the rest of this section, here $\mu$ is treated as the unknown state $\mathcal{X}$, and $\tau$ as the unknown parameter $\theta$.

**Gibbs sampler implementation:** The conditional distribution of $\mu$ is given by

$$\begin{aligned} p(\mu \mid \mu', \sigma_\mu^2, \tau, \mathcal{Y}) &\propto p(\mathcal{Y}|\mu, \tau) p(\mu \mid \mu', \sigma_\mu^2) \\ &= \mathcal{N}_\mu(\mu_0', \sigma_{\mu 0}^2), \end{aligned} \tag{2.73}$$

where

$$\mu_0' = \sigma_{\mu 0}^2 \left[ \bar{y} N \tau + \frac{\mu'}{\sigma_\mu^2} \right] \qquad \sigma_{\mu 0}^2 = \frac{1}{N\tau + \frac{1}{\sigma_\mu^2}}, \tag{2.74}$$

and $\bar{y}$ is the sample mean. Similarly, the conditional distribution of $\tau$ is given by

$$\begin{aligned} p(\tau \mid \alpha_\tau, \beta_\tau, \mu, \mathcal{Y}) &\propto p(\mathcal{Y}|\mu, \tau) p(\tau \mid \alpha_\tau, \beta_\tau) \\ &= \mathcal{G}a_\tau(\alpha_{\tau,0}, \beta_{\tau,0}), \end{aligned} \tag{2.75}$$

where

$$\alpha_{\tau,0} = \alpha_\tau + \frac{N}{2}, \qquad \beta_{\tau,0} = \beta_\tau + \sum_{j=1}^{N} \frac{(y_j - \mu)^2}{2}. \tag{2.76}$$

The Gibbs sampler is hence easily implemented by sampling $\mu^{(i+1)}$ from $p(\mu \mid \mu', \sigma_\mu^2, \tau^{(i)}, \mathcal{Y})$ and $\tau^{(i+1)}$ from $p(\tau \mid \alpha_\tau, \beta_\tau, \mu^{(i+1)}, \mathcal{Y})$. The set of samples $\{\mu^{(1..N)}\}$ $\{\tau^{(1..N)}\}$ may then be used to approximate the true posterior distributions.

**VBEM implementation:** The VBEM approximation for this problem is given as

$$\tilde{p}(\mu, \tau|\mathcal{Y}) \approx \tilde{p}(\mu)\tilde{p}(\tau). \tag{2.77}$$

From Theorem 2.3, the variational posterior at the $(i+1)^{th}$ iteration, $\tilde{p}(\mu)^{(i+1)}$, is given by

$$\begin{aligned} \tilde{p}(\mu)^{(i+1)} &\propto p(\mu|\mu', \sigma_u^2) \exp(\mathbb{E}_{\tilde{p}(\tau)^{(i)}}[\ln p(\mathcal{Y}|\mu, \tau)]) \\ &= \mathcal{N}_\mu(\mu_0', \sigma_{\mu 0}^2), \end{aligned} \tag{2.78}$$

where

$$\mu_0' = \sigma_{\mu 0}^2 \left[ \bar{y} N \mathbb{E}_{\tilde{p}(\tau)^{(i)}}[\tau] + \frac{\mu'}{\sigma_\mu^2} \right], \qquad \sigma_{\mu 0}^2 = \frac{1}{N \mathbb{E}_{\tilde{p}(\tau)^{(i)}}[\tau] + \frac{1}{\sigma_\mu^2}}. \tag{2.79}$$

Similarly, the variational posterior $\tilde{p}(\tau)^{(i+1)}$ is given by

$$\begin{aligned} \tilde{p}(\tau)^{(i+1)} &\propto p(\tau|\alpha_\tau, \beta_\tau) \exp(\mathbb{E}_{\tilde{p}(\mu)^{(i+1)}}[\ln p(\mathcal{Y}|\mu, \tau)]) \\ &= \mathcal{G}a_\tau(\alpha_{\tau,0}, \beta_{\tau,0}), \end{aligned} \tag{2.80}$$

where

$$\alpha_{\tau,0} = \alpha_\tau + \frac{N}{2}, \qquad \beta_{\tau,0} = \beta_\tau + \sum_{j=1}^{N} \frac{\mathbb{E}_{\tilde{p}(\mu)^{(i+1)}}[(y_j - \mu)^2]}{2}. \tag{2.81}$$

Since the prior distributions (2.71) are in the same parametric family as the posterior distributions (2.78) and (2.80), they are termed conjugate priors. The selection of an appropriate form for the prior distributions is an important step in VB, so that the distributional updates across iterations may be carried out analytically.

**MAP-EM implementation:** From (2.50), the E-step at the $(i+1)^{th}$ iteration is given by

$$\begin{aligned}\tilde{p}(\mu)^{(i+1)} &\propto p(\mathcal{Y}|\mu, \tau^{(i)})p(\mu|\mu', \sigma_\mu^2) \\ &= \mathcal{N}_\mu(\mu_0', \sigma_{\mu 0}^2),\end{aligned} \tag{2.82}$$

where

$$\mu_0' = \sigma_{\mu 0}^2 \left[\bar{y} N \tau^{(i)} + \frac{\mu'}{\sigma_\mu^2}\right], \qquad \sigma_{\mu 0}^2 = \frac{1}{N\tau^{(i)} + \frac{1}{\sigma_\mu^2}}. \tag{2.83}$$

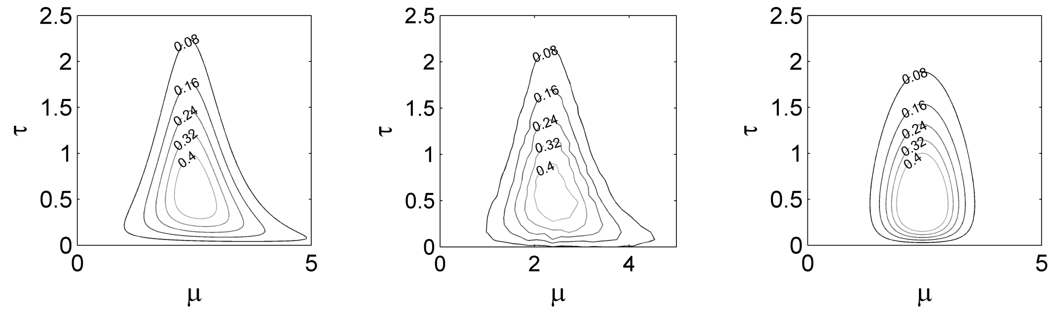
From (2.59), the M-step is given by

$$\begin{aligned}\tau^{(i+1)} &= \arg\max_\tau \left[\mathbb{E}_{\tilde{p}(\mu)^{(i+1)}}[\ln p(\mathcal{Y}|\mu,\tau)] + \ln p(\tau)\right] \\ &= \frac{\alpha_\tau - 1 + \frac{N}{2}}{\beta_\tau + \frac{1}{2}\sum_{j=1}^N \mathbb{E}_{\tilde{p}(\mu)^{(i+1)}}[(y_j - \mu)^2]}.\end{aligned} \tag{2.84}$$

Although the similarity between all the update equations across the three methods is very apparent, it is the small differences which distinguish their respective behaviours. In particular, when comparing VBEM and EM it is seen that the EM algorithm in (2.84) assigns the *mode* of the posterior Gamma distribution $((\alpha_{\tau,0} - 1)/\beta_{\tau,0})$ to $\tau^{(i+1)}$. The VBEM algorithm instead uses the *mean* $(\alpha_{\tau,0}/\beta_{\tau,0})$ throughout its update of $\tilde{p}(\mu)$ in (2.79). Comparing the VBEM updates (2.79) and (2.81) with the distributions sampled from by the Gibbs sampler (2.74) and (2.76), it is seen that the variational posteriors are in fact the conditionals averaged over the unknown variables. This averaging is why VB is commonly associated with *mean field* approximation methods[8] originating from the statistical physics literature [156].

A simulation was set up with the true parameters $\mu = 2$ and $\tau = 1$ and with $N = 4$.

[8] Mean field methods *fully* factorise the posterior distribution [155, 129, Chapter 2] and can thus be considered as a special case of VB.



(a) (b) (c)

Figure 2.2: (a) True joint posterior $p(\mu, \tau|\mathcal{Y})$ and estimated joint posterior distributions $\tilde{p}(\mu, \tau|\mathcal{Y})$ using (b) a Gibbs sampler and (c) the VBEM algorithm. In the latter case the VB approximation is $\tilde{p}(\mu, \tau|\mathcal{Y}) \approx \tilde{p}(\mu)\tilde{p}(\tau)$.

The hyperparameters (assumed to be known) were set to $\mu' = 6, \sigma_\mu^2 = 10, \alpha_\tau = 0.2$ and $\beta_\tau = 0.1$ The true posterior (2.72) is clearly non-separable in $\tau$ and $\mu$, indicating that in this case the product distribution (normal × gamma) of the VBEM yields a different functional form for the joint posterior distribution. This is evident in Figure 2.2 where it is also seen that the Gibbs sampler with 100,000 samples (with 10,000 removed to account for burn-in) accurately estimates the required distribution. However the joint variational posterior is the best possible fit of a normal × gamma distribution to the true posterior and for most practical purposes its approximation may be considered good enough.

The individual variational posteriors are also different from the true posteriors (which for this problem may not be found analytically) and those given by the Gibbs sampler. Yet, from Figure 2.3 it is seen that they are very representative. Of interest is the position of the mode of the variational posterior $\tilde{p}(\tau)$, which is seen to lie in the direction of the skewness of the true distribution, i.e. closer to the mean of the true posterior. This is again a result of the averaging and is a common observation with unimodal distributions (see Section 4.1.2). The mean and variance of the posterior distributions as given by the Gibbs sampler and VB algorithm are given in Table 2.1. For both unknown parameters, the VB algorithm is seen to under-estimate the variance. This is a common problem with VB methods [157] (see Section 3.4 for the implications in spatiotemporal systems).

Table 2.1: Comparison between the Gibbs and VBEM posterior distributions.

| | mean($\mu$) | var($\mu$) | mean($\tau$) | var($\tau$) |
|---|---|---|---|---|
| Gibbs | 2.56 | 0.65 | 0.83 | 0.4 |
| VBEM | 2.45 | 0.29 | 0.82 | 0.3 |

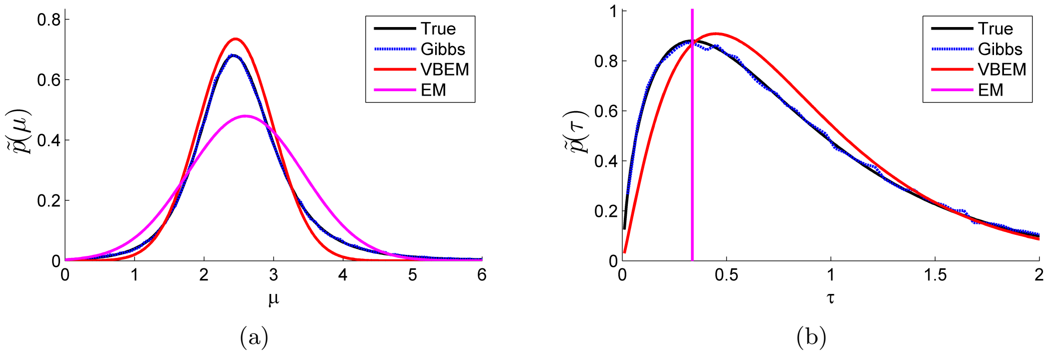


Figure 2.3: The true and estimated posteriors for (a) $\mu$ and (b) $\tau$ using the Gibbs sampler, VB algorithm and EM algorithm.

In Figure 2.3 it is seen that the MAP-EM algorithm accurately sets the maximiser $\tau^*$ to the mode of the true posterior as expected. However, the estimation of $\tilde{p}(\mu) = p(\mu|\mathcal{Y}, \tau^*)$ is clearly very different from the true posterior. Therefore if the EM algorithm is used as a joint estimator (as is commonly the case in spatiotemporal systems [e.g. 113]) the price of ML estimation of $\boldsymbol{\theta}$ may indeed be the distributional inaccuracy of the latent data. Since this thesis is particularly interested with the hidden states, VB is seen to be a plausible alternative to EM for the application at hand, particularly because the added computational cost when compared to EM is minimal.[9] For this problem the EM and VBEM algorithms both only required 100$\mu$s on a standard PC for 10 iterations. The Gibbs sampler on the other hand required 45s for 100,000 samples, the bottleneck being obviously the sampling process (despite the fact that the sampling distributions are conventional), a step omitted in the other two methods.

## 2.4 Conclusion

This chapter has presented a brief review on:

- Spatiotemporal models which have found considerable use in the application domain, namely the CML, the IDE, the PDE and their stochastic extensions. Here the SPDE was highlighted as the model class which satisfies all the objectives set out in Section 1.3.2.

- Essential properties and methods associated with SPDEs. This includes the formal

[9] Although EM and VBEM are both deterministic approximation approaches with the same computational complexity, VBEM algorithms tend to require more instruction evaluations than corresponding EM algorithms.

definition of Brownian motion and the Wiener process, properties of the abstract $Q$-Wiener process and a brief review on discretisation methods employed in practice.

- State-space models in the context of spatiotemporal modelling and associated inference methodologies, namely the EM algorithm, the VBEM algorithm and Gibbs sampling.

In doing so this chapter has introduced the concepts which will be recurring throughout the thesis. These notions will be used to obtain a state-space model from an underlying SPDE through the use of theoretical properties of the mathematical construct. Joint field-parameter estimation methods will then be used to estimate unknown quantities in the model. Chapter 3 will put this overview in context and establish the basic framework for linear SPDEs by expanding on the methods highlighted in Sections 2.3.2-2.3.4 for their inference from continuous data; Chapter 4 will then consider the even more interesting case when the methodology is adapted to cater for observations which come in the form of events rather than continuous readings.

# Chapter 3

# Field-parameter estimation from continuous observations

In Section 2.1 it was seen why the SPDE, when compared to other dynamic model classes, is ideally suited for spatiotemporal system representation for the purposes of this work. However, (S)PDEs describe infinite dimensional fields evolving on a continuous spatiotemporal domain. Measurements, on the other hand, are frequently taken at isolated locations, at discrete-time intervals and in noisy conditions. Estimation of the field and governing parameters from such potentially uninformative data is highly challenging.

Although the estimation problem has been studied extensively for the deterministic case [e.g. 46, 158], little has been done in the stochastic case [93]. Moreover, the solutions proposed to date for the stochastic case are mostly not in line with the objectives of Section 1.3.2. For instance, in [159] Sallberg used an infinite dimensional Kalman filter in conjunction with multiple model adaptive estimation to find the thermal conductivity of the stochastic heat equation. This approach requires the expansion of the semigroup of linear operators associated with the generator $\mathcal{A}$, something only possible with simple problems such as the stochastic heat equation *with spatially invariant parameters*. Wikle in [49] introduced a fully Bayesian approach which however requires the SPDE to be discretised on a regular mesh. In addition, the class of models considered in his work are in principle deterministic; the noise was added ad-hoc to cater for discretisation error and lacks physical meaning.

A notable exception is the work of Solo [93] were the Galerkin method was employed to bring the SPDE into finite dimensional form before application of the EM algorithm. The approach is an excellent starting point for the considerations in this thesis since it is seen to allow for spatially heterogeneous systems whilst maintaining a continuous-space representation. However, the estimation method proposed in [93] does not provide a means for obtaining uncertainty measures over the parameters, which could be vital for control purposes. Neither does the work provide a means for basis function selection which is a required task when employing Galerkin methods. Finally, it does not provide any simulation studies which, as will be seen, are pivotal in demonstrating advantages and limitations of the proposed methodology. It is in this light that this chapter proposes a solution to the SPDE estimation problem by consolidating ideas presented in [93].

The first step in the considered approach is a rigorous application of the Galerkin method to obtain a reduced model representation for SPDEs in Section 3.1. This is facilitated by the treatment of the additive space-time noise term using theory outlined in Section 2.2, coupled with basis function selection methods using frequency analysis. The resulting state-space model maintains a continuous-space representation of the spatiotemporal field. This representation will prove necessary for using continuous-space data in the point process treatment of Chapters 4-6.

The state-space model lends itself easily to the second step, estimation. A vast range of methods, from ML estimation augmented with parameter uncertainty measures, to approximate Bayesian techniques, to fully Bayesian methods are made possible. The potential use of algorithms lying in these three categories are shown through the implementation of the augmented EM algorithm, VBEM algorithm and a Gibbs sampler, with complete details given in Section 3.2. The section thus presents a set of new, practical identification algorithms for the SPDE state-space representation, all of which are able to provide parameter uncertainty measures. The developed methods are benchmarked against each other in Section 3.3 on a spatiotemporal case study. The VBEM method is seen to be orders of magnitude faster than the Gibbs sampler and to give results in line with the latter method.

A key result (which reaches beyond the systems considered here) made possible by studying the behaviour of the different mechanisms is that the VBEM algorithm may provide highly overconfident estimates in these systems. The effect, discussed in Section 3.4, is studied in relation to the quality of the available data and a remedy for it is given using the augmented EM class of inference mechanisms.

## 3.1 State-space representation of SPDEs

In this section the Galerkin method, described in Section 2.2.1, will be used to reduce the SPDE into a finite dimensional representation. In order to conform with the observation process, which is frequently, by nature, discrete in time, the reduced model will then be discretised using standard Euler methods; an explicit scheme is intentionally employed to maintain a 'linear in the parameters' representation. The resulting systems of equations are then coupled with the discrete-time observation process to obtain a state-space representation of the SPDE.

### 3.1.1 The Galerkin-Euler approximation

As discussed in Section 2.2.1, the first step of the Galerkin method considers the expansion of the stochastic field into a sum of basis functions weighted by some unknown random variables (or weights); subsequently the inner product of the resulting evolution equation is taken with respect to a set of test functions, which for the Galerkin method is chosen to be the same as the set of basis functions. An Euler step is then applied to obtain a discrete-time representation of the finite dimensional system. The following theorem formalises the procedure for obtaining the discrete-time finite dimensional representation of the SPDE.

**Theorem 3.1 (Discrete-time finite dimensional representation of SPDEs)** *Let $\mathcal{A} : H \rightarrow H$ describe a linear operator and $W(t) = W(\cdot, t) \in H$. Using the Galerkin approximation method in conjunction with the explicit Euler scheme, the infinite dimensional equation*

$$\begin{aligned} dz(t) &= \mathcal{A}z(t)dt + \sigma_w dW(t), \\ z(0) &= z_0, \end{aligned} \tag{3.1}$$

*may be represented as an evolution equation of the form*

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}(\boldsymbol{\vartheta})\boldsymbol{x}_k + \boldsymbol{w}_k, \tag{3.2}$$

*where $\boldsymbol{x}_k \in \mathbb{R}^n, \boldsymbol{A}(\boldsymbol{\vartheta}) \in \mathbb{R}^{n\times n}$ and $\boldsymbol{w}_k \in \mathbb{R}^n$ with $\mathbb{E}[\boldsymbol{w}_k] = \boldsymbol{0}$ and cov$[\boldsymbol{w}_k] = \boldsymbol{\Sigma}_w \in \mathbb{R}^{n\times n}$.*

*Proof.* The function $z(t) \in H$ may be written as a (possibly) infinite sum in $\phi_i$, where $\mathcal{B} = \{\phi_i\}$ forms a set of basis functions in $H$

$$z(t) = \sum_{i=1}^{\infty} x_i(t)\phi_i. \tag{3.3}$$

Denote the $n$-dimensional subspace $H_n = \text{span}\{\phi_i; 1 \leq i \leq n\}$. It is then possible to truncate the sum in (3.3) to obtain an approximation of $z(t)$ given by

$$z(t) \approx \sum_{i=1}^{n} x_i(t)\phi_i. \tag{3.4}$$

Choose $\chi_j, 1 \leq j \leq n$ as suitable sufficiently regular test functions which vanish on $\partial\mathcal{O}$ and take the inner product of the resulting SPDE with respect to each $\chi_j$ so that

$$\langle \mathrm{d}z(t), \chi_j \rangle \approx \sum_{i=1}^{n} \mathrm{d}x_i(t) \langle \phi_i, \chi_j \rangle, \tag{3.5}$$

and

$$\langle \mathcal{A}(\boldsymbol{\vartheta}) z(t) \mathrm{d}t, \chi_j \rangle \approx \sum_{i=1}^{n} x_i(t) \langle \mathcal{A}(\boldsymbol{\vartheta}) \phi_i, \chi_j \rangle \mathrm{d}t, \tag{3.6}$$

for $1 \leq j \leq n$. In the Galerkin method $\chi_j = \phi_j$. The additive noise increments have a projection $\langle \mathrm{d}W(t), \phi_j \rangle$ which by Theorem 2.1 are the increments of a real-valued stochastic process. The approximated covariance $\boldsymbol{Q}\mathrm{d}t : \mathbb{R}^n \to \mathbb{R}^n$ of the increments of the projected $\mathrm{d}W(t)$ is then given by

$$\boldsymbol{Q}\mathrm{d}t = \left[ \mathbb{E}[\langle \mathrm{d}W(t), \phi_i \rangle \langle \mathrm{d}W(t), \phi_j \rangle] \right]_{i,j=1}^{n}. \tag{3.7}$$

Again, by Theorem 2.1, the matrix $\boldsymbol{Q}$ is given by

$$\boldsymbol{Q} = [q_{i,j}]_{i,j=1}^{n}, q_{i,j} = \langle Q\phi_i, \phi_j \rangle. \tag{3.8}$$

Setting $\boldsymbol{\Psi}_x = [\langle \phi_i, \phi_j \rangle]_{i,j=1}^{n}$ and $\boldsymbol{\Psi}_A(\boldsymbol{\vartheta}) = [\langle \mathcal{A}(\boldsymbol{\vartheta}) \phi_i, \phi_j \rangle]_{i,j=1}^{n}$

$$\boldsymbol{\Psi}_x \mathrm{d}\boldsymbol{x}(t) = \boldsymbol{\Psi}_A(\boldsymbol{\vartheta}) \boldsymbol{x}(t) \mathrm{d}t + \sigma_w \mathrm{d}\boldsymbol{\beta}(t), \tag{3.9}$$

where $\boldsymbol{x}(t) = [x_1(t) \dots x_n(t)]^T$ and $\mathrm{d}\boldsymbol{\beta}(t) = [\mathrm{d}\beta_1(t) \dots \mathrm{d}\beta_n(t)]^T$ is the $n$-dimensional projected noise with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{Q}\mathrm{d}t$. Equation (3.9) is a linear set of coupled SDEs. Now, for a discretisation interval $\Delta_t$, denoting $\boldsymbol{x}_k := \boldsymbol{x}(k\Delta_t)$ and $\boldsymbol{\beta}_k := \boldsymbol{\beta}(k\Delta_t)$, an explicit Euler scheme yields a discrete-time approximate system of equations given by

$$\boldsymbol{\Psi}_x (\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) = \boldsymbol{\Psi}_A(\boldsymbol{\vartheta}) \boldsymbol{x}_k \Delta_t + \sigma_w \Delta_{\boldsymbol{\beta}_k}, \tag{3.10}$$

where $\Delta_{\boldsymbol{\beta}_k} = \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k$ has covariance $\Delta_t \boldsymbol{Q}$. Therefore

$$\boldsymbol{x}_{k+1} = (\boldsymbol{I} + \Delta_t \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{\Psi}_A(\boldsymbol{\vartheta})) \boldsymbol{x}_k + \sigma_w \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \Delta_{\boldsymbol{\beta}_k}, \tag{3.11}$$

or

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}(\boldsymbol{\vartheta}) \boldsymbol{x}_k + \boldsymbol{w}_k, \tag{3.12}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{A}(\boldsymbol{\vartheta}) = (\boldsymbol{I} + \Delta_t \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{\Psi}_A(\boldsymbol{\vartheta}))$ and $\boldsymbol{w}_k = \sigma_w \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \Delta_{\boldsymbol{\beta}_k}$ is a Gaussian random variable distributed according to $\boldsymbol{w}_k \sim \mathcal{N}_{\boldsymbol{w}_k}(\mathbf{0}, \boldsymbol{\Sigma}_w)$. Since $\boldsymbol{\Psi}_{\boldsymbol{x}}$ is symmetric, the covariance matrix of $\boldsymbol{w}_k$ is given by

$$\boldsymbol{\Sigma}_w = \sigma_w^2 \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \mathbb{E}[\Delta_{\beta_k} \Delta_{\beta_k}^T] \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} = \sigma_w^2 \Delta_t \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{Q} \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1}. \tag{3.13}$$

■

From Section 2.2.3, setting $Q = I$ gives the covariance of the increments of the projected standard cylindrical Wiener process. In the case of space-time white noise $\boldsymbol{Q} = \boldsymbol{\Psi}_{\boldsymbol{x}}$ and the covariance of the state-space evolution equation becomes $\boldsymbol{\Sigma}_w = \sigma_w^2 \Delta_t \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1}$.

To put Theorem 3.1 into perspective, consider the simulation of the one-dimension stochastic diffusion equation using the Galerkin decomposition. In this case (3.1) on $\mathcal{O} \subset \mathbb{R}$ has operator

$$\mathcal{A}(\cdot) = \frac{\partial}{\partial s}\left(D(s)\frac{\partial}{\partial s}\right)(\cdot). \tag{3.14}$$

The spatially varying diffusion coefficient $D(s)$ is assumed to be of polynomial class so that $D(s) = \vartheta_1 + \vartheta_1 s + \cdots + \vartheta_d s^d$ where $\vartheta_1 \ldots \vartheta_d$ are parameters assumed to be known for simulation purposes. The spatial differential operator may then be re-written as

$$\begin{aligned} \mathcal{A}(\cdot) &= (\vartheta_2 + 2\vartheta_3 s + \ldots (d-1)\vartheta_d s^{d-2})\frac{\partial}{\partial s}(\cdot) \\ &\quad +(\vartheta_1 + \vartheta_2 s + \ldots \vartheta_d s^{d-1})\frac{\partial^2}{\partial s^2}(\cdot). \end{aligned} \tag{3.15}$$

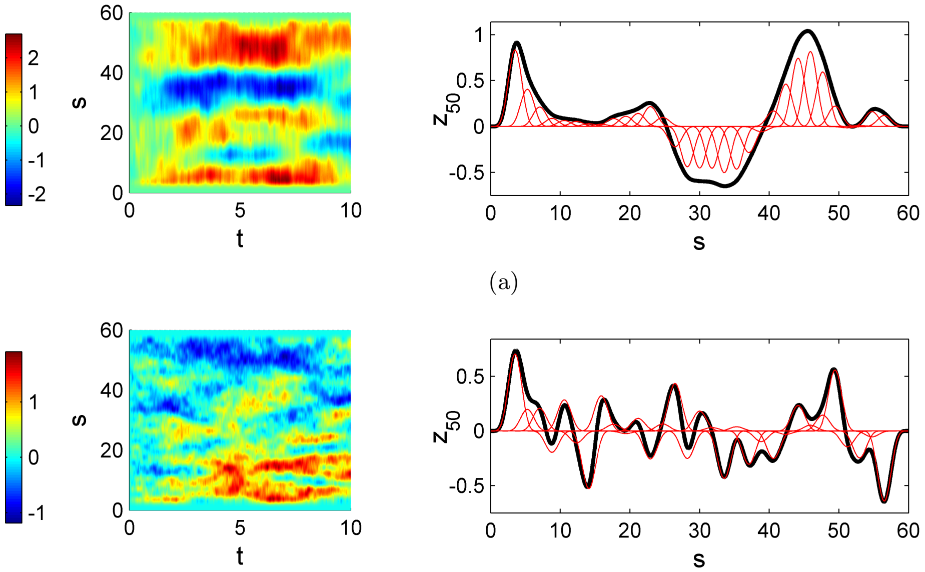
Applying Theorem 3.1 and relating the operator expansion (3.15) to (2.21), the matrix $\boldsymbol{\Psi}_A(\boldsymbol{\vartheta})$ is represented as $\boldsymbol{\Psi}_A(\boldsymbol{\vartheta}) = \sum_{i=1}^{d} \vartheta_i \boldsymbol{V}_i$ where in this case

$$\begin{aligned} \boldsymbol{V}_1 &= \int_{\mathcal{O}} \boldsymbol{\phi}(s)\frac{\partial^2 \boldsymbol{\phi}(s)^T}{\partial s^2} \mathrm{d}\mathcal{O}, \\ \boldsymbol{V}_i\,|_{i>1} &= \int_{\mathcal{O}} \boldsymbol{\phi}(s)\frac{\partial^2 \boldsymbol{\phi}(s)^T}{\partial s^2} s^{i-1} + (i-1)\boldsymbol{\phi}(s)\frac{\partial \boldsymbol{\phi}(s)^T}{\partial s} s^{i-2} d\mathcal{O}, \end{aligned} \tag{3.16}$$

and $\boldsymbol{\phi}(s) = [\phi_1(s) \ldots \phi_n(s)]^T$ is the vector of basis functions.

It follows that for a given set of basis functions, the matrices $\boldsymbol{\Psi}_A$ and $\boldsymbol{\Psi}_{\boldsymbol{x}}$ and hence $\boldsymbol{A}(\boldsymbol{\vartheta})$ are known. To find the noise covariance, $\boldsymbol{\Sigma}_w$, the quantity $\boldsymbol{Q}$ is required. This may, in turn, be found from (3.8) if $Q$ is known. Once these quantities are determined (3.12) may be forward simulated in time. The spatial field reconstruction at any time point is then given through (3.4).

A simulation using this discrete-time finite dimensional representation of the stochastic diffusion equation was set up with $\Delta_t = 0.02$, $\mathcal{O} = [0, 60]$, $D(s) = s - 0.01s + 0.001s^2$,



(a)

(b)

Figure 3.1: (a, left) A single instantiation of the stochastic diffusion equation with spatially varying diffusion coefficient, zero initial and zero boundary conditions and (a, right) the field (black) together with its comprising basis functions (red) at t = 1 for $k_Q(s) = \exp(-s^2/20)$. (b) Same as (a) with $k_Q(s) = \exp(-s^2)$.

$\sigma_w = 0.5$ and $Qu = \int_{\mathcal{O}} k_Q(s-r)u(r)\mathrm{d}r$. For this example $n = 31$ local basis functions of the form (3.20, Section 3.1.3) were chosen with $\tau = 4\pi/7$ and equally spaced throughout the spatial domain. To also show the effect of the covariance operator $k_Q$, two forms were chosen: (i) $k_Q(s) = \exp(-s^2/20)$ and (ii) $k_Q(s) = \exp(-s^2)$. Snapshots of the simulated fields, together with the individual components at $k = 50$ are shown in Figure 3.1 from where it is seen that case (i) produces a smoother field than (ii). In turn, the basis functions are able to reproduce the frequency content in the field; the problem of how to ensure this is possible from continuous data is tackled in Section 3.1.3 using Fourier methods. Finally, both fields are seen to satisfy the boundary conditions exactly as required.

### 3.1.2 The observation process

In this chapter data is assumed to be collected by $m$ sensors at distinct locations throughout the spatial domain. Each of the sensor characteristics is described through linear functions $c(\cdot)$ which define the area over which the aggregated reading is taken. These linear functions are usually assumed to be known. When the space is large in comparison to the sensor range $c(\cdot) = \delta(\cdot)$; this is usually the case with spatiotemporal systems studied on a geographical scale, such as oceanography studies with underwater vehicles [24], weather monitoring with unmanned aerial vehicles (UAV) [160, Chapter 1] and pollution monitoring with weather stations [12]. In other cases, such as neuroimaging [74], the spatial scope of the sensors cannot be ignored. Typical shape functions include the Gaussian radial basis functions (GRBF) [76], the box function [54] or the smoothed box function to maintain certain regularity assumptions and avoid the Gibbs phenomenon during numerical reconstruction [161]. In general the precise shape of $c(\cdot)$ has little effect on the performance of the algorithm when its frequency response greatly exceeds that of the underlying field, which is also in some sense a requirement in order to ensure system identifiability.

The observation process commonly associated with these types of problem may be integrated with finite dimensional reduction mechanisms to construct a linear state-space model amenable for joint state-parameter inference using methods described in Section 2.3. Consider a spatiotemporal field $z(t)$ and let $\mathcal{C}_i(t) : H \to \mathbb{R}, i = 1 \dots m$ denote the output operator of the $i^{th}$ sensor which takes an average of the underlying field at regular time intervals $\Delta_t$. Define $z(k\Delta_t) := z_k$, $\mathcal{C}_{k,i} := \mathcal{C}_i(k\Delta_t)$, $v_{k,i} \sim \mathcal{N}_{v_{k,i}}(0, \sigma_v^2)$ and the output reading of the $i^{th}$ sensor as $y_i(k\Delta_t) := y_{k,i}$. Then the sensor output readings corrupted by white noise at the $k^{th}$ time interval is related to the field as $y_{k,i} = \mathcal{C}_{k,i} z_k + v_{k,i}$. The ensemble reading is given as

$$\boldsymbol{y}_k = \left[ \int_{\mathcal{O}} c(\boldsymbol{s}; \boldsymbol{s}_{i,k}^c) z_k(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} \right]_{i=1}^{m} + \boldsymbol{v}_k, \tag{3.17}$$

where $\boldsymbol{s}_{i,k}^c \in \mathbb{R}^m$ denotes the position of the $i^{th}$ sensor at time $t = k\Delta_t$, and $\boldsymbol{v}_k \in \mathbb{R}^m$ is additive white Gaussian noise with zero mean and covariance $\boldsymbol{\Sigma}_v = \sigma_v^2 \boldsymbol{I}$.

Considering the expansion of $z_k$ on $H_n$ as in (3.3) the output equation is expressed in matrix form as

$$\boldsymbol{y}_k = \boldsymbol{C}_k \boldsymbol{x}_k + \boldsymbol{v}_k, \tag{3.18}$$

where

$$\boldsymbol{C}_k = \begin{bmatrix} \int_{\mathcal{O}} c(\boldsymbol{s}; \boldsymbol{s}_{1,k}^c) \boldsymbol{\phi}^T(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} \\ \int_{\mathcal{O}} c(\boldsymbol{s}; \boldsymbol{s}_{2,k}^c) \boldsymbol{\phi}^T(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} \\ \vdots \\ \int_{\mathcal{O}} c(\boldsymbol{s}; \boldsymbol{s}_{m,k}^c) \boldsymbol{\phi}^T(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} \end{bmatrix}. \tag{3.19}$$

If the sensors are point sensors, it follows that $\boldsymbol{C}_k^T = [\boldsymbol{\phi}(\boldsymbol{s}_{1,k}^c) \dots \boldsymbol{\phi}(\boldsymbol{s}_{m,k}^c)]$, where $\boldsymbol{\phi}(s_{j,k}^c)$ is the vector of contributions of the basis functions for the $j^{th}$ sensor located at the position $\boldsymbol{s}_j^c$ at sample number $k$. If further the sensors are static, $\boldsymbol{C}_k^T = \boldsymbol{C}^T = [\boldsymbol{\phi}(\boldsymbol{s}_1^c) \dots \boldsymbol{\phi}(\boldsymbol{s}_m^c)]$.

As illustrated above, the observation process (3.18) has been extensively employed in various application settings. It should be noted, however, that in practice it may need to be altered to account for small-scale or microscale variations which arise due to the truncation (3.4) or unmodelled effects (for instance by appropriate structuring of $\boldsymbol{\Sigma}_v$). See [162, 163] for relevant discussions.

### 3.1.3 Basis function placement from continuous observations

In the simulation example in Section 3.1.1, basis functions were used to represent an SPDE for simulation purposes. However, when modelling the SPDE for *estimation* purposes, a suitable set of basis functions is not available a priori. Moreover, what set to elicit when employing the observation process (3.17) has not been extensively studied in the context of SPDEs. For these reasons, this section discusses basis function selection and proposes a principled engineering approach adopted from the neural network literature as a way forward.

In the Galerkin method (and the method of moments in general), the only requirements are that the basis set $\mathcal{B}$ is a linearly independent set which can satisfy the approximation of (3.4) with reasonable accuracy [99, Section 1.3]. In practice the choice of $\mathcal{B}$ may be affected by the ease of evaluation of the matrix elements $\boldsymbol{Q}$, $\boldsymbol{\Psi}_{\boldsymbol{x}}$ and $\boldsymbol{\Psi}_A$ and the maximum feasible size $n$ for matrix storage and operation (such as inversion). The chosen set need not be orthogonal nor orthonormal.

A vast range of continuous-space functions satisfy these criteria and may be used to form $\mathcal{B}$. A common choice in function approximation, the Fourier series, is not wholly appropriate for spatiotemporal systems which, in general, may exhibit considerable heterogeneity in spatial behaviour (in practice this would result in an unreasonably large $n$ for accurate reconstruction). The same argument applies to sets with functions exhibiting a wide scope, such as Jacobi polynomials or, similarly, Zernike polynomials on the unit disk [30]. The hereogeneity may be better catered for by the widely used empirical orthogonal functions [162, 163] which are proven to minimise the variance of the truncation error cause by (3.4). However their construction is cumbersome when observations are not taken at regular spatial intervals and moreover are known to not adequately model the dynamics of the process under study [164, Section 7.1.3].

In this thesis, the construction of a set comprising functions of local scope, such as Epanechnikov kernels or B-splines [165] is hence suggested. This allows for direct manipulation to cater for heterogeneity (see, for example, [22] and Figure 6.5 on pg. 177) and are generally easy to construct. In the context of spatiotemporal field reconstruction GRBFs have gained particular interest recently [22, 23, 73, 166, Section 5.4] as they allow analytic solutions to some otherwise intensive computations and are proven to have universal approximation properties [167]. Yet, they may not be used in the presence of Dirichlet boundary conditions since they are not of compact support and hence do not vanish on $\partial\mathcal{O}$.

As a result, the use of the following radial basis function (RBF), termed the *local GRBF*, is proposed for use throughout this work [168]:

$$\phi(\boldsymbol{s}) = \begin{cases} \frac{(2\pi - ||\tau \boldsymbol{s}||)(1+(\cos||\tau \boldsymbol{s}||)/2)+\frac{3}{2}\sin(||\tau \boldsymbol{s}||)}{3\pi}, & ||\tau \boldsymbol{s}|| < 2\pi, \\ 0, & \text{otherwise}, \end{cases} \tag{3.20}$$

for $\tau > 0$ and where $||\cdot||$ denotes the usual Euclidean distance on $\mathcal{O}$. The local GRBF closely resembles the usual isotropic GRBF with $\phi(\boldsymbol{s}) = \exp(-\tau^2||\boldsymbol{s}||^2/2\pi)$, but is of compact support. This function was first introduced in [169] as a truncated covariance function in order to force independence between points in space which are sufficiently far away from each other.

When compared to the use of GRBFs, this basis function enhances numerical conditioning when computing the Gram matrix[1] $\boldsymbol{\Psi_x}$, and also improves computational speed-up when computing inverses as a result of increased matrix sparseness. A key advantage is that as a result of the close resemblance, with only a slight loss in accuracy, one may use standard analytic solutions associated with the usual GRBF when carrying out the Galerkin decomposition. More importantly, results for function reconstruction using standard GRBFs may be readily applied for placing them within the spatial domain.

The principal idea proposed here for basis function placement is to use the results of Sanner and Slotine in [170]. First, the local GRBF functions are arranged regularly in the spatial domain such that any subset of $\mathcal{O}$ may be adequately represented in the reconstruction. The resolution of the grid has to be small enough to avoid aliasing by satisfying Shannon's sampling criterion. In particular, if the centres of the basis functions, denoted $\{\boldsymbol{\zeta}_i\}_{i=1}^n$, are set equal to the sequence of vectors describing a regular lattice of edge length $\Delta_s$ in $\mathcal{O}$, then it is required that

$$\Delta_s < \frac{1}{2\nu_c} = \frac{1}{2\alpha_0\nu_c}, \tag{3.21}$$

where $\nu_c$ denotes the support of the signal's frequency content, assumed to be the same in each of the components of $\boldsymbol{s}$, and $\alpha_0$ is an oversampling parameter.

Second, it is required to find the range, or extent, of the local functions which in turn determines the range of frequencies they can represent. Consider the case when the basis

[1] a Hermitian matrix of inner products

functions are GRBFs with functional form

$$\phi(s) = \exp\left(-\frac{s^2}{2\sigma_b^2}\right). \tag{3.22}$$

The Fourier transform of the GRBF is yet another GRBF (in the frequency domain) given as

$$\phi(\nu) = \mathcal{F}\{\phi(s)\} = \sqrt{2\pi\sigma_b^2}\exp(-2\pi^2\sigma_b^2\nu^2), \tag{3.23}$$

so that the variances in the spatial and frequency domain are related through the mappings [170]

$$\sigma_\nu^2 \leftarrow \frac{1}{4\pi^2\sigma_b^2} \qquad \sigma_b^2 \leftarrow \frac{1}{4\pi^2\sigma_\nu^2}. \tag{3.24}$$

This relationship is very convenient as it may be used to select the width of the GRBFs in the spatial domain based on its width in the frequency domain.

The range of frequencies which can be represented by the basis functions has to exceed that of the field for adequate reconstruction. To this end Sanner and Slotine suggest the following relationship

$$\sigma_\nu = \frac{1}{\sqrt{2}}\nu_c. \tag{3.25}$$

Given $\sigma_\nu$, relationship (3.24) can then be used to find the width of the desired GRBF in $\mathcal{O}$. By simple substitution of (3.25) in (3.24)

$$\sigma_b = \sqrt{\frac{1}{2\nu_c^2\pi^2}}. \tag{3.26}$$

To find the required parameter $\tau$ for local GRBFs one may apply the relation

$$\tau = \sqrt{\pi}/\sigma_b = \sqrt{2\nu_c^2\pi^3}. \tag{3.27}$$

The resulting basis is hence a set of local GRBFs with parameter $\tau$ placed in the spatial domain centred on the coordinates $\{\boldsymbol{\zeta}_i\}_{i=1}^n$.

The only quantity required for basis function placement is simply $\nu_c$ (the support of the signal frequency content) and an oversampling parameter $\alpha_0$. In the case of continuous data observations, $\nu_c$ can be estimated using spatial Fourier analysis, as shown in the case study at the end of this chapter. A similar approach employing frequency analysis is also possible for basis function placement when dealing with point process observations; a detailed approach for this special case will be discussed in detail in Chapter 4.

Irrespective of the observation process, throughout this thesis it will be assumed that the spatial frequency response of the field is temporally invariant (so that a temporally averaged Fourier transform may be used to compute $\nu_c$). While this assumption may be a restriction in some cases, for example in time-varying systems (see Chapter 5), it greatly facilitates algorithm development and improves computational efficiency. In particular the system dimensionality is ensured to be temporally invariant and the matrices $\boldsymbol{\Psi}_{\boldsymbol{x}}$, $\boldsymbol{V}$ and $\boldsymbol{Q}$ may be computed before inference is carried out.

## 3.2 SPDE field and parameter estimation

The previous section outlined a strategy for converting SPDEs with operators of the form (2.21) observed with noisy sensors into a conventional state-space form. As a result of such a representation, well-established and also more recent statistical estimation techniques ranging from ML to approximate and full Bayesian techniques may be used for joint field-parameter inference. To demonstrate this, corresponding estimation methods, discussed in Section 2.3, are developed here in the context of SPDEs. First, the EM algorithm of [93] is derived but further augmented to provide parameter uncertainty measures. Then the VBEM algorithm is outlined and finally a full MCMC method is applied; the prime use of the latter will be to verify the other two methods. These methods are also applied to similar systems with point process observations in Chapter 4.

Note that the set of equations (3.2) and (3.18) describe a LDS with the unknown states $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=0}^K$, $\boldsymbol{x}_0 \sim \mathcal{N}_{\boldsymbol{x}_0}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, observed data $\mathcal{Y} = \{\boldsymbol{y}_i\}_{i=1}^K$, unknown parameters[2] $\boldsymbol{\theta}_g = (\boldsymbol{\vartheta}, \sigma_w^{-2}, \sigma_v^{-2})$ and, following the convention in [129, Chapter 5], unknown hyperparameters describing the distribution of the initial state $\boldsymbol{\theta}_h = (\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0^{-1})$. The distinction between hyperparameters and parameters is only relevant in the Bayesian context and the unknown quantities may be collected together as $\boldsymbol{\theta} = (\boldsymbol{\theta}_g, \boldsymbol{\theta}_h)$. Throughout it is assumed that the operator $Q$ is known a priori. The problem of SPDE identification may be solved by estimating the states governing the field statistics $\mathcal{X}$ and the unknown quantities $\boldsymbol{\theta}$ given the observed data. A graphical representation of the model is given in Figure 3.2.

### 3.2.1 Augmented EM methods for SPDEs

**E-Step:** Unlike the simple static system considered in the case study of Section 2.3.5, the data collected from dynamic systems may not be considered to be i.i.d. as the underlying state sequence forms a Markov chain (see Figure 3.2). The Markov structure

[2] Precisions instead of variances are included in $\boldsymbol{\theta}$ to facilitate derivations later on.

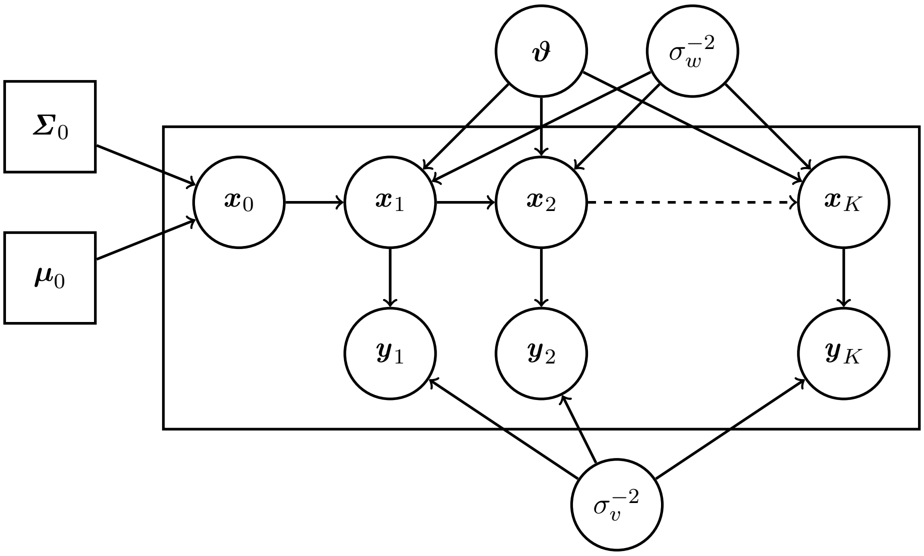


Figure 3.2: Graphical representation of the reduced state-space model for SPDEs with unknown parameters. The box encloses the model of Figure 2.3.

yields a joint likelihood for the reduced SPDE model given by

$$p(\mathcal{X}, \mathcal{Y} \mid \boldsymbol{\theta}) = p(\boldsymbol{x}_0) \prod_{k=1}^{K} p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1}, \boldsymbol{\vartheta}, \sigma_w^{-2}) \prod_{k=1}^{K} p(\boldsymbol{y}_k \mid \boldsymbol{x}_k, \sigma_v^{-2}). \tag{3.28}$$

In Section 2.3.2 it was shown that for the E-step at the $i^{th}$ iteration, the distribution $\tilde{p}(\mathcal{X})^{(i+1)} = p(\mathcal{X}|\mathcal{Y}, \boldsymbol{\theta}^{(i)})$ needs to be evaluated for carrying out the expectation of (3.28). As stated in Section 2.3.1, for the LDS it is well known that relevant quantities from this distribution are given by the smoother developed by RTS, given in the context of SPDEs in Algorithm A.1 in the appendix.

The smoother provides posterior distributions of the states at each $k$, which are normally distributed of the form $\mathcal{N}_{\boldsymbol{x}_k}(\hat{\boldsymbol{x}}_{k|K}, \boldsymbol{\Sigma}_{k|K})$. For the linear Gaussian model considered, the quantities required for the M-step are the first order moment $\mathbb{E}[\boldsymbol{x}_k] = \hat{\boldsymbol{x}}_{k|K}$, the second order moment $\mathbb{E}[\boldsymbol{x}_k \boldsymbol{x}_k^T]$ and the cross second-order moment $\mathbb{E}[\boldsymbol{x}_k \boldsymbol{x}_{k-1}^T]$. The former second moment may be found through the state smoothed covariance matrix $\boldsymbol{\Sigma}_{k|K}$ as obtained directly from the smoother. Computation of the latter quantity requires the computation of the cross covariance matrix $\boldsymbol{M}_{k|K}$ [25] as shown in the last *for loop* of Algorithm A.1.

**M-Step:** Having computed the required statistics on $\mathcal{X}$ under the parameter vector $\boldsymbol{\theta}^{(i)}$, it is now possible to compute the required expectation of (3.28) which in turn is required to find $\boldsymbol{\theta}^{(i+1)}$ in the M-step (2.53). From (3.28) the expected log likelihood is given by

$$\begin{aligned}\mathbb{E}_{\tilde{p}(\mathcal{X})}[\ln p(\mathcal{X},\mathcal{Y}|\boldsymbol{\theta})] = \mathbb{E}_{\tilde{p}(\mathcal{X})}\Bigg[&\ln p(\boldsymbol{x}_0|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) + \sum_{k=1}^{K}\ln p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1},\boldsymbol{\vartheta},\sigma_w^2)\\ &+\sum_{k=1}^{K}\ln p(\boldsymbol{y}_k \mid \boldsymbol{x}_k,\sigma_v^2)\Bigg].\end{aligned} \tag{3.29}$$

Consider the identity

$$|\sigma^2\boldsymbol{\Sigma}| = \sigma^{2n}|\boldsymbol{\Sigma}|, \qquad \boldsymbol{\Sigma}\in\mathbb{R}^{n\times n}. \tag{3.30}$$

Then, under the Gaussianity assumptions (ignoring constants relative to $\boldsymbol{\theta}$) (3.29) is given as[3]

$$\begin{aligned}\mathbb{E}_{\tilde{p}(\mathcal{X})}[\ln p(\mathcal{X},\mathcal{Y}|\boldsymbol{\theta})] = \mathbb{E}_{\tilde{p}(\mathcal{X})}\Bigg[&-\frac{1}{2}\ln|\boldsymbol{\Sigma}_0| - \frac{1}{2}(\boldsymbol{x}_0-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}_0-\boldsymbol{\mu}_0)\\ &+\frac{Kn}{2}\ln\sigma_w^{-2} - \frac{1}{2}\sum_{k=1}^{K}\sigma_w^{-2}(\boldsymbol{x}_k-\boldsymbol{A}(\boldsymbol{\vartheta})\boldsymbol{x}_{k-1})^T\widetilde{\boldsymbol{Q}}^{-1}(\boldsymbol{x}_k-\boldsymbol{A}(\boldsymbol{\vartheta})\boldsymbol{x}_{k-1})\\ &+\frac{Km}{2}\ln\sigma_v^{-2} - \frac{1}{2}\sum_{k=1}^{K}\sigma_v^{-2}(\boldsymbol{y}_k-\boldsymbol{C}_k\boldsymbol{x}_k)^T(\boldsymbol{y}_k-\boldsymbol{C}_k\boldsymbol{x}_k)\Bigg],\end{aligned} \tag{3.31}$$

where from Theorem 3.1,

$$\widetilde{\boldsymbol{Q}} = \Delta_t\,\boldsymbol{\Psi}_{\boldsymbol{x}}^{-1}\boldsymbol{Q}\,\boldsymbol{\Psi}_{\boldsymbol{x}}^{-1}, \tag{3.32}$$

and $\boldsymbol{A}(\boldsymbol{\vartheta}) = \left(\boldsymbol{I}+\Delta_t\,\boldsymbol{\Psi}_{\boldsymbol{x}}^{-1}\sum_{i=1}^{d}\boldsymbol{V}_i\vartheta_i\right)$. The required updates for the parameters are given in the following theorem.

**Theorem 3.2** *Maximisation of (3.31) at the $(i+1)^{th}$ iteration yields the ML estimates $\boldsymbol{\theta}^{(i+1)} = [\boldsymbol{\vartheta}^{(i+1)^T},\sigma_w^{-2^{(i+1)}},\sigma_v^{-2^{(i+1)}},\boldsymbol{\mu}_0^{(i+1)},\boldsymbol{\Sigma}_0^{-1^{(i+1)}}]^T$ given by*

$$\boldsymbol{\mu}_0^{(i+1)} = \hat{\boldsymbol{x}}_{0|K}, \tag{3.33}$$

$$\boldsymbol{\Sigma}_0^{-1^{(i+1)}} = \boldsymbol{\Sigma}_{0|K}^{-1}, \tag{3.34}$$

$$\boldsymbol{\vartheta}^{(i+1)} = \boldsymbol{\Upsilon}^{-1}\boldsymbol{\upsilon}, \tag{3.35}$$

$$\sigma_w^{-2^{(i+1)}} = \frac{Kn}{\boldsymbol{\Pi}_w}, \tag{3.36}$$

$$\sigma_v^{-2^{(i+1)}} = \frac{Km}{\boldsymbol{\Pi}_v}, \tag{3.37}$$

[3] If $\boldsymbol{x}$ is normally distributed with mean $\boldsymbol{\mu}\in\mathbb{R}^n$ and covariance $\boldsymbol{\Sigma}\in\mathbb{R}^{n\times n}$, then its probability density function is given by $p(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$.

*where*

$$\boldsymbol{\upsilon} = \Delta_t[tr(\boldsymbol{V}_i^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} (\boldsymbol{\Gamma}_{1:K} - \boldsymbol{\Lambda}_{0:K-1}))]_{i=1}^d, \tag{3.38}$$

$$\boldsymbol{\Upsilon} = \Delta_t^2[tr(\boldsymbol{V}_i^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{V}_j \boldsymbol{\Lambda}_{0:K-1})]_{i,j=1}^d, \tag{3.39}$$

$$\boldsymbol{\Pi}_w = tr(\widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Lambda}_{1:K} - 2\boldsymbol{A}(\boldsymbol{\vartheta}^{(i+1)})^T \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Gamma}_{1:K} + \boldsymbol{A}(\boldsymbol{\vartheta}^{(i+1)})^T \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{A}(\boldsymbol{\vartheta}^{(i+1)}) \boldsymbol{\Lambda}_{0:K-1}), \tag{3.40}$$

$$\boldsymbol{\Pi}_v = \sum_{k=1}^{K} \left( \boldsymbol{y}_k^T \boldsymbol{y}_k - 2\boldsymbol{y}_k^T \boldsymbol{C}_k \hat{\boldsymbol{x}}_{k|K} + tr(\boldsymbol{C}_k^T \boldsymbol{C}_k \boldsymbol{\Lambda}_k) \right), \tag{3.41}$$

$tr(\cdot)$ *denotes the matrix trace operator and where*

$$\boldsymbol{\Gamma}_{k_1:k_2} = \sum_{k=k_1}^{k_2} \boldsymbol{\Gamma}_k = \sum_{k=k_1}^{k_2} \mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}[\boldsymbol{x}_k \boldsymbol{x}_{k-1}^T] = \sum_{k=k_1}^{k_2} [\hat{\boldsymbol{x}}_{k|K} \hat{\boldsymbol{x}}_{k-1|K}^T + \boldsymbol{M}_{k|K}], \tag{3.42}$$

$$\boldsymbol{\Lambda}_{k_1:k_2} = \sum_{k=k_1}^{k_2} \boldsymbol{\Lambda}_k = \sum_{k=k_1}^{k_2} \mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}[\boldsymbol{x}_k \boldsymbol{x}_k^T] = \sum_{k=k_1}^{k_2} [\hat{\boldsymbol{x}}_{k|K} \hat{\boldsymbol{x}}_{k|K}^T + \boldsymbol{\Sigma}_{k|K}]. \tag{3.43}$$

*Proof.* See Appendix B.1. ■

Joint field-parameter estimation for the SPDE is thus obtained by iteratively smoothing the state sequence and maximising to obtain parameter estimates. The EM algorithm for SPDEs is summarised in Algorithm 3.1.

**Algorithm 3.1** The EM algorithm for SPDEs

```
Input: Data set $\mathcal{Y}$, parameters $\{\boldsymbol{C}_k\}_{k=1}^K$, $\widetilde{\boldsymbol{Q}}$, $\{\boldsymbol{V}_i\}_{i=1}^d$, $\boldsymbol{\Psi}_{\boldsymbol{x}}$ and initial parameter estimates $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_g^{(0)}, \boldsymbol{\theta}_h^{(0)}) = (\boldsymbol{\vartheta}^{(0)}, \sigma_w^{-2^{(0)}}, \sigma_v^{-2^{(0)}}, \boldsymbol{\mu}_0^{(0)}, \boldsymbol{\Sigma}_0^{-1^{(0)}})$.

$i = 0$
do
    Run Algorithm A.1 with $\boldsymbol{\theta} = \boldsymbol{\theta}^{(i)}$          E-step
    Evaluate $\boldsymbol{\theta}^{(i+1)}$ from (3.33)-(3.37)          M-step
    $i = i + 1$
until $||\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}^{(i-1)}|| < \epsilon_1$

Output: $\{\hat{\boldsymbol{x}}_{k|K}, \boldsymbol{\Sigma}_{k|K}\}_{k=0}^K, \boldsymbol{\theta}^{(i)}$.
```

In spatiotemporal systems the parameter $\boldsymbol{\vartheta}$ is of key interest as it describes the dynamics of the system. A measure of parameter uncertainty on $\boldsymbol{\vartheta}$ may prove to be vital in decision making or in strategic sensor planning when studying SPDEs. However, the conventional EM algorithm, as described above and in [93], does not provide confidence intervals for the ML estimates. Nonetheless there exist ways of evaluating or estimating the information matrices associated with the estimated dynamics.

### Parameter uncertainty measures with EM

The first method employed for covariance matrix evaluation with EM was developed by Louis [171] and exploits the use of the *missing information principle.* Soon after, Meng and Rubin [172] developed the supplemented EM (SEM) algorithm which uses the trajectory of the EM algorithm and the complete-data Hessian to estimate the incomplete-data Hessian. The advantage of SEM is that it obviates the requirement for computing score statistics. More recently, Duan and Fulop [173] developed a numerical approach to compute the expected incomplete-data Hessian using expected complete-data scores which are directly available from the EM algorithm.

*Louis' method:* Let the *observed-data likelihood* (denoted by an $o$ superscript) of the unknown parameter dynamics $\boldsymbol{\vartheta}$ be given as $p(\mathcal{Y}|\boldsymbol{\vartheta})$ and the *complete-data likelihood* (denoted by a $c$ superscript) be given as $p(\mathcal{X},\mathcal{Y}|\boldsymbol{\vartheta})$. Then the score statistics and the Hessians are given by

$$\boldsymbol{s}^o(\mathcal{Y},\boldsymbol{\vartheta}) = \frac{\partial \ln p(\mathcal{Y}|\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}, \qquad \boldsymbol{H}^o(\mathcal{Y},\boldsymbol{\vartheta}) = \frac{\partial^2 \ln p(\mathcal{Y}|\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}\partial \boldsymbol{\vartheta}^T}, \tag{3.44}$$

$$\boldsymbol{s}^c(\mathcal{X},\mathcal{Y},\boldsymbol{\vartheta}) = \frac{\partial \ln p(\mathcal{X},\mathcal{Y}|\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}, \qquad \boldsymbol{H}^c(\mathcal{X},\mathcal{Y},\boldsymbol{\vartheta}) = \frac{\partial^2 \ln p(\mathcal{X},\mathcal{Y}|\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}\partial \boldsymbol{\vartheta}^T}. \tag{3.45}$$

Given a parameter estimate $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^*$, it is common practice to define $\boldsymbol{I}^o(\mathcal{Y},\boldsymbol{\vartheta})_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^*} = -\boldsymbol{H}^o(\mathcal{Y},\boldsymbol{\vartheta})_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^*}$ (rather than the Fisher information matrix) as the observed information of the ML estimate [171, 174] such that $\boldsymbol{I}^{o^{-1}}(\mathcal{Y},\boldsymbol{\vartheta})_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^*}$ is the covariance of the estimate. This quantity may be found using the following theorem.

**Theorem 3.3 (Missing information principle [171])** *The observed data information matrix $\boldsymbol{I}^o(\mathcal{Y},\boldsymbol{\vartheta}) = -\boldsymbol{H}^o(\mathcal{Y},\boldsymbol{\vartheta})$ is given by*

$$\boldsymbol{I}^o(\mathcal{Y},\boldsymbol{\vartheta}) = \mathcal{I}^c(\mathcal{Y},\boldsymbol{\vartheta}) - \mathcal{I}^m(\mathcal{Y},\boldsymbol{\vartheta}), \tag{3.46}$$

*where $\mathcal{I}^c$ is the expected complete-data information matrix and $\mathcal{I}^m$ is the expected missing-data information matrix given by*

$$\mathcal{I}^c(\mathcal{Y},\boldsymbol{\vartheta}) = -\mathbb{E}_{p(\mathcal{X}|\mathcal{Y},\boldsymbol{\vartheta})}[\boldsymbol{H}^c(\mathcal{X},\mathcal{Y},\boldsymbol{\vartheta})], \tag{3.47}$$

$$\mathcal{I}^m(\mathcal{Y},\boldsymbol{\vartheta}) = \mathbb{E}_{p(\mathcal{X}|\mathcal{Y},\boldsymbol{\vartheta})}[\boldsymbol{s}^c(\mathcal{X},\mathcal{Y},\boldsymbol{\vartheta})\boldsymbol{s}^{c^T}(\mathcal{X},\mathcal{Y},\boldsymbol{\vartheta})] - \boldsymbol{s}^o(\mathcal{Y},\boldsymbol{\vartheta})\boldsymbol{s}^{o^T}(\mathcal{Y},\boldsymbol{\vartheta}). \tag{3.48}$$

*Moreover, on convergence of the EM algorithm to a stationary point $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^*$,*

$\boldsymbol{s}^o(\mathcal{Y},\boldsymbol{\vartheta})|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^*} = \mathbf{0}$ *so that*

$$\begin{aligned}\boldsymbol{I}^o(\mathcal{Y},\boldsymbol{\vartheta})_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^*} = &-\mathbb{E}_{p(\mathcal{X}|\mathcal{Y},\boldsymbol{\vartheta}^*)}[\boldsymbol{H}^c(\mathcal{X},\mathcal{Y},\boldsymbol{\vartheta})]_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^*}\\ &-\mathbb{E}_{p(\mathcal{X}|\mathcal{Y},\boldsymbol{\vartheta}^*)}[\boldsymbol{s}^c(\mathcal{X},\mathcal{Y},\boldsymbol{\vartheta})\boldsymbol{s}^{c^T}(\mathcal{X},\mathcal{Y},\boldsymbol{\vartheta})]_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^*}.\end{aligned} \tag{3.49}$$

Since both the matrices in (3.49) are given in terms of the complete-data statistics, in the case of i.i.d. data they can easily be found using the EM algorithm. However, when correlations in $(\mathcal{X},\mathcal{Y})$ exist, as in the case of dynamic systems, the second term in (3.49) becomes very complicated as quantities such as $\mathbb{E}_{\tilde{p}(\mathcal{X})}[\boldsymbol{x}_i\boldsymbol{x}_j^T]$ for all $i,j=1\ldots K$ need to be evaluated. The methods described next avoid this problem and are hence preferable for information estimation in dynamic systems.

*SEM:* The aim of the SEM is to provide a stable estimate of the asymptotic covariance matrix without the use of the score statistics so that the problem observed with the missing information principle in Theorem 3.3 when associated with dynamic systems is avoided. The method is based on [15] which, when considering a single-parameter system $(\boldsymbol{\vartheta}=\vartheta)$, relates the asymptotic variance based on the observed data $\sigma_{\vartheta}^{2^o}$ with that based on the complete-data $\sigma_{\vartheta}^{2^c}$ as follows

$$\sigma_{\vartheta}^{2^o} = \sigma_{\vartheta}^{2^c}/(1-r), \tag{3.50}$$

or

$$\sigma_{\vartheta}^{2^o} = \sigma_{\vartheta}^{2^c} + \Delta\sigma_{\vartheta}^{2^c}, \tag{3.51}$$

where $\Delta\sigma_{\vartheta}^{2^c} = [r/(1-r)]\sigma_{\vartheta}^{2^c}$ and $r$ is the rate of convergence of the EM algorithm. From (3.50) it is immediately apparent that the higher the rate of convergence,[4] the higher the discrepancy between the observed-data information and the complete-data information. The SEM algorithm for when $\boldsymbol{\vartheta}$ is a vector is given by

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^o &= \boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^c + \Delta\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^c\\ &= \boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^c + (\boldsymbol{I}-\boldsymbol{R}(\boldsymbol{\vartheta}^*))^{-1}\boldsymbol{R}(\boldsymbol{\vartheta}^*)\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^c,\end{aligned} \tag{3.52}$$

where $\boldsymbol{R}(\boldsymbol{\vartheta}^*)$ is the rate of convergence matrix in the limit of the optimal estimate $\boldsymbol{\vartheta}^*$. The diagonal elements of $\Delta\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^c$ effectively give the increase in the asymptotic variance of $\boldsymbol{\vartheta}^*$ as a result of missing information. Thus to obtain the asymptotic variance, in what is referred to as the *S-step*, one simply needs to compute $\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^c$ and $\boldsymbol{R}(\boldsymbol{\vartheta}^*)$. The first

[4] Note that here the rate of convergence corresponds to the gradient of the trajectory of the EM parameter estimates ‘close’ to the ML estimate. It is hence large when the EM algorithm requires several iterations to converge to the ML estimate.

term is the inverse of the expected complete-data information matrix (3.47). The matrix associated with the information relating to dynamics is then given by

$$\mathcal{I}_c(\boldsymbol{\vartheta}) = \sigma_w^{-2^{(i)}} \boldsymbol{\Upsilon}^{(i)} \mid_{i \to \infty} . \tag{3.53}$$

Therefore

$$\boldsymbol{\Sigma}^c_{\boldsymbol{\vartheta}} = \mathcal{I}_c^{-1}(\boldsymbol{\vartheta}) = [\sigma_w^{-2^{(i)}} \boldsymbol{\Upsilon}^{(i)}]^{-1} \mid_{i \to \infty} . \tag{3.54}$$

Details of obtaining $R(\boldsymbol{\vartheta}^*)$ through numerical differentiation are given in [131, Section 4.5.2]. Care must be taken when using estimates for $\boldsymbol{\vartheta}_j^{(i)}$ which have converged for the required computation as a result of numerical conditioning. In practice it proved hard to obtain consistent results from SEM in the case studies considered, thus warranting investigation into the third, rather less known method for parameter uncertainty estimation used with EM.

*Duan's method:* The possibility of non-negative definiteness in the empirical computation of $\boldsymbol{H}^c(\mathcal{X}, \mathcal{Y}, \boldsymbol{\vartheta})$ was the prime motivation of Duan and Fulop to develop a stable estimator of the expected (Fisher) incomplete-data matrix given by

$$\mathcal{I}^o(\mathcal{Y}, \boldsymbol{\vartheta}^*) = -\mathbb{E}_{p(\mathcal{Y}|\boldsymbol{\theta})}[\boldsymbol{H}^o(\mathcal{Y}, \boldsymbol{\theta})]. \tag{3.55}$$

The expected incomplete-data matrix defines the variance of the ML estimate in relation to the true parameter $\boldsymbol{\vartheta}_0$ since

$$(\boldsymbol{\vartheta}^* - \boldsymbol{\vartheta}_0) \sim \mathcal{N}_{\boldsymbol{\vartheta}^*}(\mathbf{0}, \mathcal{I}^{o^{-1}}), \tag{3.56}$$

and thus constitutes an uncertainty measure of the ML estimate. The estimator makes use of the numerical scheme by Newey and West [175] and a result by Louis [171] which relates the observed-data score to smoothed individual scores so that

$$\mathcal{I}^o(\mathcal{Y}, \boldsymbol{\vartheta}^*) \approx \boldsymbol{I}_0(\mathcal{Y}, \boldsymbol{\vartheta}^*) + \sum_{j=1}^{l} w_j [\boldsymbol{I}_j(\mathcal{Y}, \boldsymbol{\vartheta}^*) + \boldsymbol{I}_j^T(\mathcal{Y}, \boldsymbol{\vartheta}^*)], \quad l \geq 1, \tag{3.57}$$

where

$$w_j = 1 - \frac{j}{l+1}, \tag{3.58}$$

and

$$\boldsymbol{I}_j(\mathcal{Y}, \boldsymbol{\vartheta}^*) = \sum_{i=1}^{K-j} \mathbb{E}_{\tilde{p}(\mathcal{X})^*}[\boldsymbol{s}_k^c(\mathcal{X}, \mathcal{Y}, \boldsymbol{\vartheta}^*)] \mathbb{E}_{\tilde{p}(\mathcal{X})^*}[\boldsymbol{s}_{k+j}^c(\mathcal{X}, \mathcal{Y}, \boldsymbol{\vartheta}^*)^T]. \tag{3.59}$$

Here $\boldsymbol{s}_k^c(\mathcal{X}, \mathcal{Y}, \boldsymbol{\vartheta}^*)$ denotes the individual score at the $k^{th}$ time point, i.e.

$$\boldsymbol{s}_k^c(\mathcal{X}, \mathcal{Y}, \boldsymbol{\vartheta}^*) = \left. \frac{\partial \ln p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k, \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} \right|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^*}, \tag{3.60}$$

where the observation process has been omitted since it is independent of $\boldsymbol{\vartheta}$. For the SPDE the expectation of these individual scores are easily computed to give

$$\mathbb{E}_{\tilde{p}(\mathcal{X})^*}[\boldsymbol{s}_k^c(\mathcal{X}, \mathcal{Y}, \boldsymbol{\vartheta}^*)] = \sigma_w^{-2^*} \boldsymbol{\upsilon}_k - \sigma_w^{-2^*} \boldsymbol{\Upsilon}_k \boldsymbol{\vartheta}^*, \tag{3.61}$$

where

$$\boldsymbol{\upsilon}_k = \Delta_t [\mathrm{tr}( \boldsymbol{V}_i^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} (\boldsymbol{\Gamma}_k - \boldsymbol{\Lambda}_{k-1}))]_{i=1}^d, \tag{3.62}$$

$$\boldsymbol{\Upsilon}_k = \Delta_t^2 [\mathrm{tr}( \boldsymbol{V}_i^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{V}_j \boldsymbol{\Lambda}_{k-1})]_{i,j=1}^d. \tag{3.63}$$

As in SEM the observed information may be evaluated at the final EM iteration. Since Duan's method does not require any further EM iterations for generating the rate matrix it is generally quicker than SEM. Moreover it is more well-behaved since the requirement for numerical differentiation to obtain the rate matrix is obviated.

### 3.2.2 VBEM estimation

As outlined in Section 2.3.3, the first step when employing VBEM methods is to choose an appropriate family of distributions with which to approximate the true posterior. It is first noted that since $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0^{-1}$ are treated as hyperparameters governing the prior distribution on $\boldsymbol{x}_0$ they may be maximised in a way similar to EM [129, Section 5.3.6] (although distributions over the hyperparameters are possible). Recalling that the parameters and hyperparameters are denoted as $\boldsymbol{\theta}_g$ and $\boldsymbol{\theta}_h$ respectively, the natural choice of the variational posteriors for the SPDE under investigation is then given as

$$\begin{aligned} p(\mathcal{X}, \boldsymbol{\theta}_g|\mathcal{Y}, \boldsymbol{\theta}_h) &= p(\mathcal{X}, \boldsymbol{\vartheta}, \sigma_w^{-2}, \sigma_v^{-2}|\mathcal{Y}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0^{-1}) \\ &\approx \tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\vartheta})\tilde{p}(\sigma_w^{-2})\tilde{p}(\sigma_v^{-2}) \\ &= \tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta}_g), \end{aligned} \tag{3.64}$$

and by symmetry in Theorem 2.3 the maximisers are given by

$$\tilde{p}(\mathcal{X}) \propto \exp\Big(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_g)}[\ln p(\mathcal{X}, \boldsymbol{\theta}, \mathcal{Y})]\Big), \tag{3.65}$$

$$\tilde{p}(\boldsymbol{\vartheta}) \propto \exp\Big(\mathbb{E}_{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta}_g)^{/\boldsymbol{\vartheta}}}[\ln p(\mathcal{X}, \boldsymbol{\theta}, \mathcal{Y})]\Big), \tag{3.66}$$

$$\tilde{p}(\sigma_w^{-2}) \propto \exp\Big(\mathbb{E}_{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta}_g)^{/\sigma_w^{-2}}}[\ln p(\mathcal{X}, \boldsymbol{\theta}, \mathcal{Y})]\Big), \tag{3.67}$$

$$\tilde{p}(\sigma_v^{-2}) \propto \exp\Big(\mathbb{E}_{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta}_g)^{/\sigma_v^{-2}}}[\ln p(\mathcal{X}, \boldsymbol{\theta}, \mathcal{Y})]\Big), \tag{3.68}$$

where $\tilde{p}(\boldsymbol{\theta}_g)^{/\bar{\boldsymbol{\theta}}}$ refers to the joint $\tilde{p}(\boldsymbol{\theta}_g)$ without the $\bar{\boldsymbol{\theta}}$ component. Hence, for instance, $\tilde{p}(\boldsymbol{\theta}_g)^{/\boldsymbol{\vartheta}} = \tilde{p}(\sigma_w^{-2})\tilde{p}(\sigma_v^{-2})$.

**VBE-step:** As a result of the observation process the set of equations (3.2) and (3.18) describe a LDS, and hence a smoother similar to that by RTS may be derived to obtain the variational posterior at the $i^{th}$ iteration $\tilde{p}(\mathcal{X})^{(i)}$. However, since the parameters are also governed by distributions (and are not point estimates as assumed by RTS) the parameter moments need to be dealt with appropriately. This has two consequences:

1. The matrix inversion lemma cannot be used to obtain the neat closed form expressions generally associated with the Kalman filter and its smoothing equivalent [129, Section 5.3.3].

2. A two-filter approach is required since the usual forward-backward smoothing method leads to expressions not amenable under expectation. As shown in Section 2.3.1 the smoothed estimate is then constructed as $\tilde{p}(\boldsymbol{x}_k|\boldsymbol{y}_{1:K}) \propto \tilde{p}(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})\tilde{p}(\boldsymbol{y}_{k+1:K}|\boldsymbol{x}_k) = \alpha(\boldsymbol{x}_k)\beta(\boldsymbol{x}_k)$ where $\alpha(\boldsymbol{x}_k)$ the forward message and $\beta(\boldsymbol{x}_k)$ the backward message.

The variational Kalman smoother was derived from a Bayesian perspective by Beal in [176]. In his derivation Beal opted to fix the state noise covariance matrix $\boldsymbol{\Sigma}_w = \boldsymbol{I}$ in order to ensure identifiability of the state-space model. In spatiotemporal systems this is not a plausible argument since $\boldsymbol{\Sigma}_w$ has a pre-determined structure and there is generally a known parameterisation of $\boldsymbol{A}(\boldsymbol{\vartheta})$ so that the model has a unique representation.

In the SPDE case the inclusion of $\boldsymbol{\Sigma}_w$ is trivial and the expectations are easy to handle as $\mathbb{E}_{\tilde{p}(\sigma_w^{-2})}[\boldsymbol{\Sigma}_w^{-1}] = \mathbb{E}_{\tilde{p}(\sigma_w^{-2})}[\sigma_w^{-2}]\widetilde{\boldsymbol{Q}}^{-1}$. However, despite being linear in $\boldsymbol{\vartheta}$, the parameterisation of $\boldsymbol{A}(\boldsymbol{\vartheta})$ introduces some complications as regards to the expectations.

In particular the computation of the following quantity is required:

$$\begin{aligned}
&\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})}[\boldsymbol{A}(\boldsymbol{\vartheta})^T \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{A}(\boldsymbol{\vartheta})] \\
&\quad = \mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})}[(\boldsymbol{I} + \Delta_t \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{\vartheta}})^T \widetilde{\boldsymbol{Q}}^{-1} (\boldsymbol{I} + \Delta_t \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{\vartheta}})] \\
&\quad = \Big( \widetilde{\boldsymbol{Q}}^{-1} + \Delta_t \mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})}[\boldsymbol{\Psi}_{\boldsymbol{\vartheta}}^T] \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} + \Delta_t \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})}[\boldsymbol{\Psi}_{\boldsymbol{\vartheta}}] \\
&\qquad + \Delta_t^2 \mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})}[\boldsymbol{\Psi}_{\boldsymbol{\vartheta}}^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{\vartheta}}] \Big). \qquad (3.69)
\end{aligned}$$

In (3.69) the second and third terms of the last equality can be computed using the expected values of $\boldsymbol{\vartheta}$. The fourth on the other hand, is computed as follows

$$\begin{aligned}
\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})}[\boldsymbol{\Psi}_{\boldsymbol{\vartheta}}^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{\vartheta}}] &= \mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})} \left[ \left( \sum_{i=1}^d \vartheta_i \boldsymbol{V}_i^T \right) \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \left( \sum_{j=1}^d \vartheta_j \boldsymbol{V}_j \right) \right] \\
&= \sum_{i=1}^d \mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})}[\vartheta_i^2] \boldsymbol{V}_i^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{V}_i \\
&\quad + 2 \sum_{i,j=1; i \neq j}^d \mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})}[\vartheta_i \vartheta_j] \boldsymbol{V}_i^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{V}_j, \qquad (3.70)
\end{aligned}$$

where the appropriate second moments are found from the parameter covariance matrix as calculated in the VBM-step. In the interest of brevity the derivation of the variational Kalman smoother applied to SPDEs is omitted and the algorithm is given in Algorithm A.2 in the appendix.

**VBM-step:** Before carrying out the VBM-step, prior distributions which are conjugate with the likelihood need to be allocated to the parameters. Since the form of the likelihood is Gaussian it follows that appropriate priors are given as

$$p(\boldsymbol{\vartheta}) = \mathcal{N}_{\boldsymbol{\vartheta}}(\hat{\boldsymbol{\vartheta}}_p, \boldsymbol{\Sigma}_{\boldsymbol{\vartheta},p}), \qquad (3.71)$$

$$p(\sigma_w^{-2}) = \mathcal{G}a_{\sigma_w^{-2}}(\alpha_{w,p}, \beta_{w,p}), \qquad (3.72)$$

$$p(\sigma_v^{-2}) = \mathcal{G}a_{\sigma_v^{-2}}(\alpha_{v,p}, \beta_{v,p}). \qquad (3.73)$$

The variational posteriors at the $(i+1)^{th}$ iteration are given by the following theorem.

**Theorem 3.4** *The maximisers (3.66)-(3.68) are given by*

$$\tilde{p}(\boldsymbol{\vartheta})^{(i+1)} = \mathcal{N}_{\boldsymbol{\vartheta}}(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}), \tag{3.74}$$

$$\tilde{p}(\sigma_w^{-2})^{(i+1)} = \mathcal{G}a_{\sigma_w^{-2}}(\alpha_w, \beta_w), \tag{3.75}$$

$$\tilde{p}(\sigma_v^{-2})^{(i+1)} = \mathcal{G}a_{\sigma_v^{-2}}(\alpha_v, \beta_v), \tag{3.76}$$

*where*

$$\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}} = \big(\mathbb{E}_{\tilde{p}(\sigma_w^{-2})^{(i)}}[\sigma_w^{-2}]\,\boldsymbol{\Upsilon} + \boldsymbol{\Sigma}_{\boldsymbol{\vartheta},p}^{-1}\big)^{-1}, \tag{3.77}$$

$$\hat{\boldsymbol{\vartheta}} = \boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}\big(\mathbb{E}_{\tilde{p}(\sigma_w^{-2})^{(i)}}[\sigma_w^{-2}]\boldsymbol{\upsilon} + \boldsymbol{\Sigma}_{\boldsymbol{\vartheta},p}^{-1}\hat{\boldsymbol{\vartheta}}_p\big), \tag{3.78}$$

$$\alpha_w = \alpha_{w,p} + \frac{Kn}{2}, \tag{3.79}$$

$$\beta_w = \beta_{w,p} + \frac{\boldsymbol{\Pi}'_w}{2}, \tag{3.80}$$

$$\alpha_v = \alpha_{v,p} + \frac{Km}{2}, \tag{3.81}$$

$$\beta_v = \beta_{v,p} + \frac{\boldsymbol{\Pi}_v}{2}, \tag{3.82}$$

*and where $\boldsymbol{\Upsilon}, \boldsymbol{\upsilon}, \boldsymbol{\Pi}_v$ are given in (3.38), (3.39) and (3.41) respectively whilst $\boldsymbol{\Pi}'_w$ is $\boldsymbol{\Pi}_w$ in (3.40) under the variational transform to give*

$$\boldsymbol{\Pi}'_w = tr(\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Lambda}_{1:K} - 2\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})^{(i+1)}}[\boldsymbol{A}(\boldsymbol{\vartheta})^T]\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Gamma}_{1:K} + \mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})^{(i+1)}}[\boldsymbol{A}(\boldsymbol{\vartheta})^T\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{A}(\boldsymbol{\vartheta})]\boldsymbol{\Lambda}_{0:K-1}). \tag{3.83}$$

*Finally, the hyperparameter updates are given as*

$$\boldsymbol{\mu}_0^{(i+1)} = \hat{\boldsymbol{x}}_{0|K}, \tag{3.84}$$

$$\boldsymbol{\Sigma}_0^{-1^{(i+1)}} = \boldsymbol{\Sigma}_{0|K}^{-1}. \tag{3.85}$$

*Proof.* See Appendix B.2. ■

The VBEM algorithm for SPDEs is summarised in Algorithm 3.2. The algorithm is terminated when the expected values of the unknown parameters stabilise. Other termination conditions are possible, such as monitoring the change in the free energy [129, Section 5.3.7].

Whilst the VBE-step differs from the E-step in the EM algorithm as it includes additional parameter moments, the VBM-step is remarkably similar to the M-step. In particular if one were to carry out MAP-EM (see Section 2.3.2), the resulting equations would be identical to (3.77)-(3.82) with expectations carried out under a delta Dirac distribution centred at the ML estimate. Another point of interest is that the parameter dynamics covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}$ in (3.77) is essentially simply the inverse of the complete-data information matrix (3.54). By the missing information principle in Theorem 3.3 this quantity is larger in some sense than the observed-data information, indicating that the VB algorithm will be overconfident in its estimation. As noted in Section 3.4 this bears considerable implications when using $\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}$ as a measure of parameter uncertainty.

**Algorithm 3.2** The VBEM algorithm for SPDEs

**Input:** Data set $\mathcal{Y}$, parameters $\{\boldsymbol{C}_k\}_{k=1}^K$, $\widetilde{\boldsymbol{Q}}$, $\{\boldsymbol{V}_i\}_{i=1}^d$, $\boldsymbol{\Psi}_{\boldsymbol{x}}$, initial parameter distributions $\tilde{p}(\boldsymbol{\theta}_g)^{(0)} = \tilde{p}(\boldsymbol{\vartheta})^{(0)}\tilde{p}(\sigma_w^2)^{(0)}\tilde{p}(\sigma_v^2)^{(0)}$ and initial hyperparameters $\boldsymbol{\mu}_0^{(0)}$, $\boldsymbol{\Sigma}_0^{(0)}$.

```
$i = 0$
do
    Run Algorithm A.2 with $\tilde{p}(\boldsymbol{\theta}_g) = \tilde{p}(\boldsymbol{\theta}_g)^{(i)}$ and $\boldsymbol{\theta}_h = \boldsymbol{\theta}_h^{(i)}$        VBE-step
    Evaluate $\tilde{p}(\boldsymbol{\theta}_g)^{(i+1)}$ from (3.74)-(3.76)        VBM-step
    Evaluate $\boldsymbol{\theta}_h^{(i+1)}$ from (3.84)-(3.85)        Hyperparameter update
    $i = i + 1$
until $||\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_g)^{(i)}}[\boldsymbol{\theta}_g] - \mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_g)^{(i-1)}}[\boldsymbol{\theta}_g]|| < \epsilon_1$
```

**Output:** $\{\hat{\boldsymbol{x}}_{k|K}, \boldsymbol{\Sigma}_{k|K}\}_{k=0}^K, \tilde{p}(\boldsymbol{\theta}_g)^{(i)}, \boldsymbol{\theta}_h^{(i)}$.

### 3.2.3 Gibbs sampling

As shown in Algorithm 2.3, the Gibbs sampler may be used to estimate the posterior distributions $p(\mathcal{X}|\mathcal{Y})$ and $p(\boldsymbol{\theta}_g|\mathcal{Y})$ by iteratively sampling from the conditionals $p(\mathcal{X}|\boldsymbol{\theta}_g^{(i)}, \boldsymbol{\theta}_h, \mathcal{Y})$ and $p(\boldsymbol{\theta}_g|\mathcal{X}^{(i+1)}, \mathcal{Y})$. For the SPDE case under study the Gibbs sampler would make use of the following distributions:

$$\begin{aligned}
\mathcal{X}^{(i+1)} &\sim p(\mathcal{X}|\boldsymbol{\vartheta}^{(i)}, \sigma_w^{-2^{(i)}}, \sigma_v^{-2^{(i)}}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0^{-1}, \mathcal{Y}), && (3.86)\\
\boldsymbol{\vartheta}^{(i+1)} &\sim p(\boldsymbol{\vartheta}|\mathcal{X}^{(i+1)}, \sigma_w^{-2^{(i)}}, \mathcal{Y}), && (3.87)\\
\sigma_w^{-2^{(i+1)}} &\sim p(\sigma_w^{-2}|\mathcal{X}^{(i+1)}, \boldsymbol{\vartheta}^{(i+1)}, \mathcal{Y}), && (3.88)\\
\sigma_v^{-2^{(i+1)}} &\sim p(\sigma_v^{-2}|\mathcal{X}^{(i+1)}, \mathcal{Y}). && (3.89)
\end{aligned}$$

Since conditional dependencies exist within the parameter vector (in particular between $\boldsymbol{\vartheta}$ and $\sigma_w^{-2}$), the Gibbs sampler described by (3.86)-(3.89) is a multi-stage Gibbs sampler which is the natural extension of the basic two-stage Gibbs sampler described in Section 2.3.4.

In [152] Carter and Kohn proposed a forward-filtering backward-sampling algorithm to generate samples from $p(\mathcal{X}|\boldsymbol{\theta}, \mathcal{Y})$ in LDSs. Their approach is seen to dominate Gibbs samplers which generate states one at a time from the conditional $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{x}_{k+1}, \boldsymbol{\theta}, \boldsymbol{y}_k)$. The algorithm makes use of the following factorisation (refer to Figures 2.3 and 3.2)

$$p(\boldsymbol{x}_{1:K}|\boldsymbol{\theta}, \boldsymbol{y}_{1:K}) = p(\boldsymbol{x}_K|\boldsymbol{\theta}, \boldsymbol{y}_{1:K}) \prod_{k=0}^{K-1} p(\boldsymbol{x}_k|\boldsymbol{x}_{k+1}, \boldsymbol{\theta}, \boldsymbol{y}_{1:k}). \tag{3.90}$$

It hence first applies a standard Kalman filter to obtain the state filter distribution $p(\boldsymbol{x}_K|\boldsymbol{\theta}, \boldsymbol{y}_{1:K})$ from which $\boldsymbol{x}_K$ is sampled and then generates $\boldsymbol{x}_k$ from $p(\boldsymbol{x}_k|\boldsymbol{x}_{k+1}, \boldsymbol{\theta}, \boldsymbol{y}_{1:k})$.

The key idea to obtain the backward sampling densities is to transform the standard forward equation

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{w}_k, \tag{3.91}$$

into

$$\tilde{\boldsymbol{x}}_{k+1} = \tilde{\boldsymbol{A}}\boldsymbol{x}_k + \tilde{\boldsymbol{w}}_k, \tag{3.92}$$

where $\tilde{\boldsymbol{w}}_k$ has a diagonal covariance matrix. Then the transformed samples at $k+1$ act as noisy observations for $\tilde{\boldsymbol{A}}\boldsymbol{x}_k$ so that the observation update step of the Kalman filter may be applied to obtain the required distribution over $\boldsymbol{x}_k$. In practice this transformation may carried out by employing an $\boldsymbol{LDL}$ decomposition [177, Chapter 1] of the state evolution covariance $\boldsymbol{\Sigma}_w$. This yields $\tilde{\boldsymbol{x}}_{k+1} = \boldsymbol{L}^{-1}\boldsymbol{x}_{k+1}$, $\tilde{\boldsymbol{A}} = \boldsymbol{L}^{-1}\boldsymbol{A}$ and $\tilde{\boldsymbol{w}}_k = \boldsymbol{L}^{-1}\boldsymbol{w}_k$ with a diagonal covariance matrix $\boldsymbol{D}$. The output of the algorithm is a sequence of sampling distributions $p(\boldsymbol{x}_k|\boldsymbol{x}_{k+1}, \boldsymbol{\theta}, \boldsymbol{y}_{1:k}), k = 0 \ldots K-1$ which are normal and hence can be easily sampled from. Details of its implementation can be found in [152, Appendix 1].

The parameter sampling distributions are obtained from (3.87)-(3.89), the closed form of which are found in the same way as that for the VBEM algorithm. These are given as

$$\boldsymbol{\vartheta}^{(i+1)} \sim \mathcal{N}_{\boldsymbol{\vartheta}}(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}), \tag{3.93}$$

$$\sigma_w^{-2^{(i+1)}} \sim \mathcal{G}a_{\sigma_w^{-2}}(\alpha_w, \beta_w), \tag{3.94}$$

$$\sigma_v^{-2^{(i+1)}} \sim \mathcal{G}a_{\sigma_v^{-2}}(\alpha_v, \beta_v), \tag{3.95}$$

where $\boldsymbol{\vartheta}^{(i+1)}$ denotes the $(i+1)^{th}$ sample of $\boldsymbol{\vartheta}$ and

$$\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}} = \big([\sigma_w^{-2^{(i)}}]\,\boldsymbol{\Upsilon}^s + \boldsymbol{\Sigma}_{\boldsymbol{\vartheta},p}^{-1}\big)^{-1}, \tag{3.96}$$

$$\hat{\boldsymbol{\vartheta}} = \boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}\big([\sigma_w^{-2^{(i)}}]\boldsymbol{\upsilon}^s + \boldsymbol{\Sigma}_{\boldsymbol{\vartheta},p}^{-1}\hat{\boldsymbol{\vartheta}}_p\big), \tag{3.97}$$

$$\alpha_w = \alpha_{w,p} + \frac{Kn}{2}, \tag{3.98}$$

$$\beta_w = \beta_{w,p} + \frac{\boldsymbol{\Pi}_w^s}{2}, \tag{3.99}$$

$$\alpha_v = \alpha_{v,p} + \frac{Km}{2} \tag{3.100}$$

$$\beta_v = \beta_{v,p} + \frac{\boldsymbol{\Pi}_v^s}{2}, \tag{3.101}$$

where $\boldsymbol{\upsilon}^s, \Upsilon^s, \boldsymbol{\Pi}_w^s$ and $\boldsymbol{\Pi}_v^s$ are the same quantities given in (3.38)-(3.41) but with the state variables replaced by the samples (with the expectation dropping out).

It is also possible to obtain sampling distributions for $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$, however to mimic the behaviour of the EM algorithm and VBEM algorithm these hyperparameters will be assigned the mean and covariance of the Markov chain corresponding to $\boldsymbol{x}_0$ on convergence (as in (3.84) and (3.85)). The resulting Gibbs sampler for the SPDE is given in Algorithm 3.3.

**Algorithm 3.3** The Gibbs sampler for SPDEs

**Input:** Data set $\mathcal{Y}$, fixed parameters $\{\boldsymbol{C}_k\}_{k=1}^K$, $\widetilde{\boldsymbol{Q}}$, $\{\boldsymbol{V}_i\}_{i=1}^d$, $\boldsymbol{\Psi}_{\boldsymbol{x}}$, initial parameter samples $\boldsymbol{\vartheta}^{(0)}, \sigma_w^{2^{(0)}}, \sigma_v^{2^{(0)}}$ and hyperparameters $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$.

```
for $i = 1$ to $N$
    for $k = 1$ to $K$
        Find $p(\boldsymbol{x}_k|\boldsymbol{\theta}_g^{(i)}, \boldsymbol{\theta}_h, \boldsymbol{y}_{1:k})$        Forward filtering (Kalman filter)
    end
    Sample $\boldsymbol{x}_K^{(i)} \sim p(\boldsymbol{x}_K|\boldsymbol{\theta}_g^{(i)}, \boldsymbol{\theta}_h, \mathcal{Y})$
    for  $k = K-1$ to $0$
        Sample $\boldsymbol{x}_k^{(i)} \sim p(\boldsymbol{x}_k|\boldsymbol{x}_{k+1}, \boldsymbol{\theta}_g^{(i)}, \boldsymbol{\theta}_h, \boldsymbol{y}_{1:k})$        Backward sampling [152]
    end
    $\mathcal{X}^{(i)} \leftarrow \{\boldsymbol{x}_k^{(i)}\}_{i=0}^K$
    Sample $\boldsymbol{\vartheta}^{(i+1)}, \sigma_w^{2^{(i+1)}}, \sigma_v^{2^{(i+1)}}$ from (3.93)-(3.95)        Parameter sampling
    $\boldsymbol{\theta}_g^{(i+1)} \leftarrow (\boldsymbol{\vartheta}^{(i+1)}, \sigma_w^{-2^{(i+1)}}, \sigma_v^{-2^{(i+1)}})$
end for
```

**Output:** $\{\mathcal{X}^{(i)}\}_{i=0}^N, \{\boldsymbol{\theta}_g^{(i)}\}_{i=0}^N$.

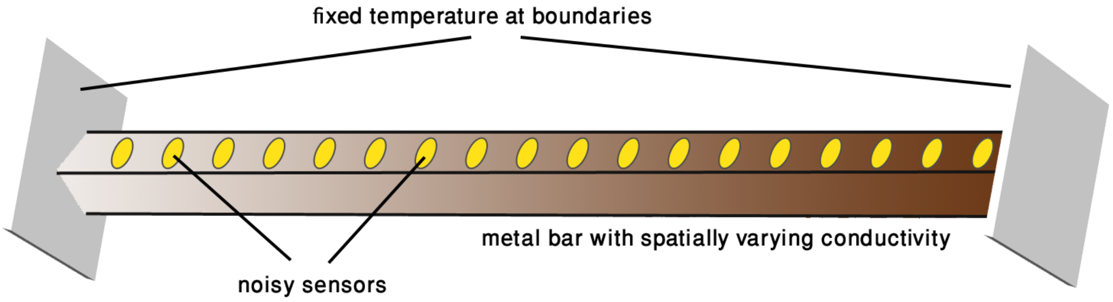

fixed temperature at boundaries
metal bar with spatially varying conductivity
noisy sensors

Figure 3.3: Schematic depicting experimental setup for the case study. The metal bar has a temperature field which is fixed at both ends (Dirichlet boundary conditions) and sensors regularly spaced along its length.

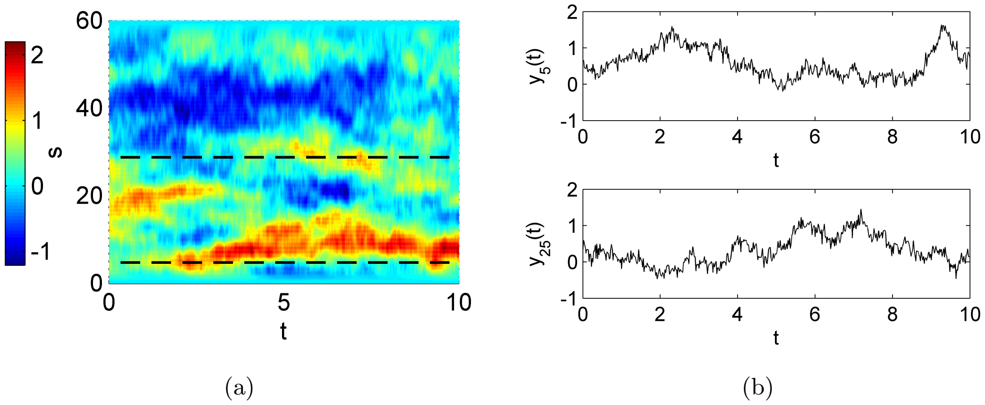
## 3.3 Case study: the stochastic diffusion equation

Consider the stochastic diffusion equation with the operator (3.14). In some cases, an analytic solution for $D(s)$ may be obtained for the deterministic PDE under steady-state conditions [178]. However, in the presence of unknown parameters, random disturbances and noisy observations the estimation problem is significantly harder and warrants the use of the methods proposed in this chapter. This stochastic equation has been used to model, for instance, the spatiotemporal behaviour of the temperature of a metal bar in the presence of quickly changing environmental conditions [179] and with fixed temperature at both ends.

### 3.3.1 Simulation setup

For this study, synthetic data was generated by considering a one-dimensional spatial domain $\mathcal{O} = [0, 60]$ with a fine mesh of $n^{sim} = 61$ basis functions of the form (3.20) with $\tau^{sim} = 4.2$.[5] The functions were equally spaced on $\mathcal{O}$ such that the Dirichlet boundary conditions were satisfied. Simulation parameters were set as $\Delta_t^{sim} = 0.02$, $\sigma_w = 0.5$, $\sigma_v = 0.1$, $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$ and $Qu = \int_{\mathcal{O}} k_Q(s-r)u(r)\mathrm{d}r$ with $k_Q(s) = \exp(-s^2/4)$. The spatially varying parameter chosen was $D(s) = 1 + 0.002s^2$ so that $\boldsymbol{\vartheta} = [1, 0, 0.002]^T$. A synthetic field was generated through forward simulation of the SPDE as described in the example of Section 3.1.1 with $K = 500$. Data was then assumed to be collected with the same sampling interval of $\Delta_t = 0.02$ from 51 point sensors equally spaced in the domain. A schematic of the simulation setup with equally spaced sensors measuring the field is shown in Figure 3.3. Two sensor readings are shown for a single instantiation in

[5] The parameters $n, \tau$ and $\Delta_t$ used in simulation need not be identical to those used in modelling. The superscript *sim* is used to differentiate between the two sets of parameters.



(a)

(b)

Figure 3.4: (a) Single instantiation of the stochastic diffusion equation with varying diffusion parameter and (b) selected output readings. Note how regions of poor conductivity ($s \in [0, 20]$) experience larger fields than regions of high conductivity ($s \in [40, 60]$).

Figure 3.4.

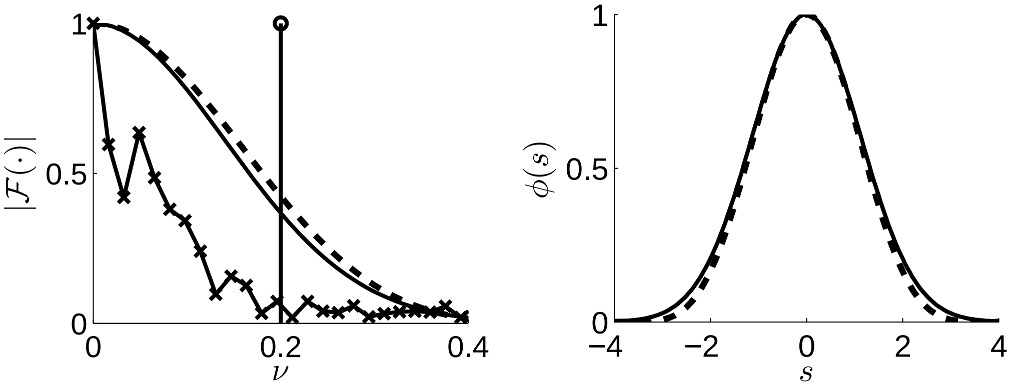
The only assumptions taken before modelling the system were that the additive disturbance is Gaussian with some known covariance operator $Q$ and that the varying parameter in space may be adequately represented as a polynomial of degree 2. The latter assumption may be further relaxed with the use of model selection methods, not considered in this work. The following sections apply the proposed methodology to estimate the field and unknown parameter vector $\boldsymbol{\theta}$ and discuss in detail the results and algorithm behaviour.

### 3.3.2 Basis function placement and estimation

Basis functions able to reconstruct the spatiotemporal field were elicited according to the guidelines of Section 3.1.3. To obtain the spatial frequency support $\nu_c$, the average spatial frequency response over the entire time horizon was found and an approximate value was then determined graphically. For the data shown in Figure 3.4, $\nu_c$ was estimated to be 0.2 cycles per unit. Recall that, given $\nu_c$, the local GRBF width is given by

$$\tau = \sqrt{2\nu_c^2\pi^3}. \tag{3.102}$$

which in this case results in a width of $\tau = 1.58$. With this value of $\tau$, 32 local GRBFs were placed regularly interspaced in $\mathcal{O}$ corresponding to an oversampling parameter of $\alpha_0 = 1.5$. The average frequency response of a single instantiation of the signal, its support and the Fourier representation of the GRBF and the local GRBF equivalent are shown in Figure 3.5. As expected the discrepancy between the GRBF and the local



(a)

(b)

Figure 3.5: (a) Average frequency response of field (marked line), estimated support of field frequency content (vertical line) and frequency response of the GRBF (solid line) and the local GRBF (dashed line). (b) Spatial representation of the GRBF (solid line) and local GRBF (dashed line).

GRBF are seen to be minimal, both in the frequency domain and in the spatial domain.

After basis function placement, the appropriate matrices $\boldsymbol{\Psi}_{\boldsymbol{x}}$ and $\{\boldsymbol{V}_i\}_{i=1}^d$ were found. This allowed application of the augmented EM algorithm, VB algorithm and Gibbs sampler to successfully estimate the unknown field and parameters and relevant uncertainties. In the Bayesian methods uninformative priors[6] were associated with all parameters under study: $\hat{\boldsymbol{\vartheta}}_p = \mathbf{0}$, $\boldsymbol{\Sigma}_{\boldsymbol{\vartheta},p} = 1 \times 10^5 \boldsymbol{I}$, $\bar{\alpha}_{w,p} = \bar{\alpha}_{v,p} = 0.1$, $\bar{\beta}_{w,p} = \bar{\beta}_{v,p} = 0.01$. The initial variational posteriors were set equal to the respective priors. In this case 40 iterations were employed for both the EM and the VBEM algorithm and $N = 10,000$ samples were drawn using the Gibbs sampler; for EM and VBEM convergence was assessed graphically. Convergence of the Gibbs sampler was confirmed using trace plots.

### 3.3.3 Results and discussion

In Figure 3.6 the parameter estimation trajectories given by the EM algorithm (ML estimate) and VBEM algorithm (mean estimate) are given. Note the initial fluctuations in the trajectory with the VBEM algorithm; these are most likely due to the large initial parameter uncertainty attributed to the parameter dynamics. On the first iteration, as a result of this uncertainty the states are estimated with high precision (a direct characteristic of the system's bilinear form), thus contributing to an over-estimation of the state precision gain $\sigma_w^{-2}$. The algorithm subsequently requires a few iterations to

[6] Where by uninformative it is implied that the characteristics of the posterior distribution are largely dominated by the data.

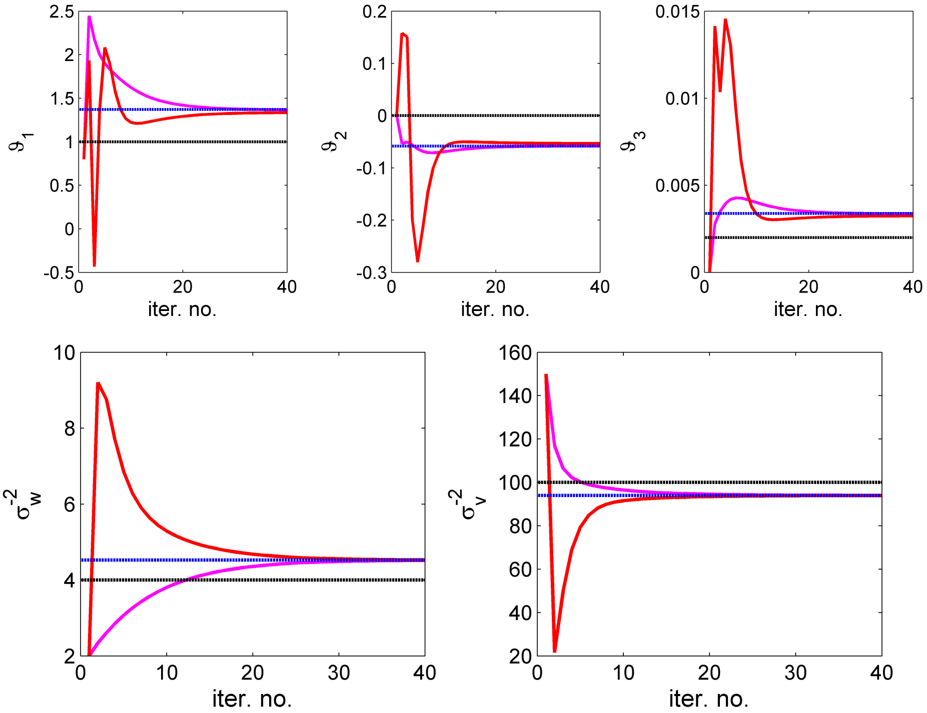
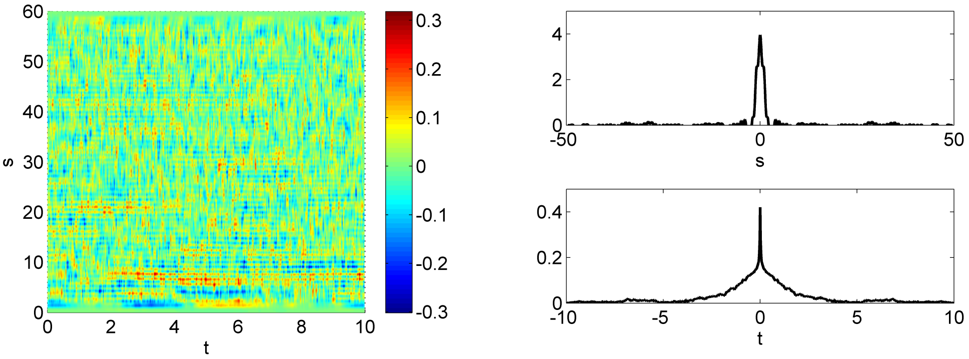


Figure 3.6: Convergence of estimates of $\boldsymbol{\vartheta}$, $\sigma_w^{-2}$ and $\sigma_v^{-2}$ to the true mean as given by the Gibbs sampler (blue). The magenta curve denotes the ML estimate of the EM algorithm whilst the red curve denotes the mean as given by the VBEM algorithm. The true value is denoted by the black line. Confidence intervals are omitted for sake of clarity.

stabilise and follow trajectories which are more similar to those given by conventional EM. These fluctuations may be limited or controlled by introducing informative priors or by initialising with a contained parameter uncertainty measure.

To check the quality of the estimated field, in Figure 3.7 the error between the true field and reconstructed field using the VBEM posterior mean is shown, together with the error autocorrelation signals in space and time. The spatial autocorrelation tests yield a response which decays quickly to zero (in relation to the width of the modelling basis functions) whilst the residuals are seen to be correlated in time, a direct consequence of the model mismatch between the simulation and modelling dimensionality. This mismatch is also seen to play a role in parameter estimation (see Section 3.3.4).

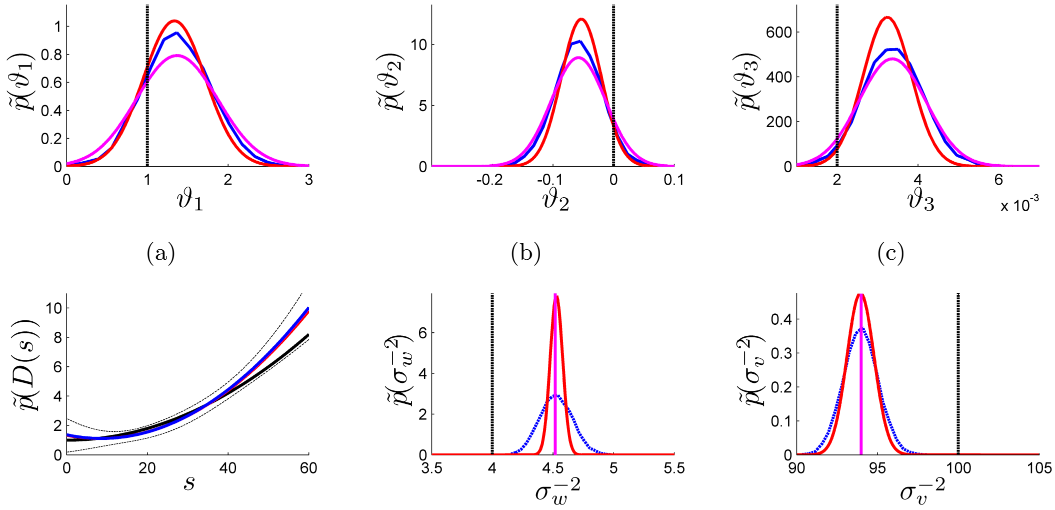
The posterior distributions of the parameter vector $\boldsymbol{\vartheta}$ are shown in Figure 3.8a-3.8c. Since $D(s)$ is a polynomial function, the spread of the parameters does not come as a surprise and indeed it is much more indicative to show the mean and uncertainty of $D(s)$ as constructed from the statistics of $\boldsymbol{\vartheta}$ (Figure 3.8d). However the distributions over the individual parameters are still shown as they further evidence the algorithms comparing



(a)

(b)

Figure 3.7: Error analysis in a simulation using a higher-order representation than that used in modelling. (a) Field error $z - \hat{z}$ where $\hat{z}$ is the reconstructed field using the smoothed states from the VBEM algorithm (EM algorithm gives similar results). (b) Average spatial autocorrelation across all time points (top) and average temporal autocorrelation across all spatial points (bottom).



(a) (b) (c)

(d) (e) (f)

Figure 3.8: Posterior distributions and ML estimates as obtained from the EM algorithm (magenta), VBEM algorithm (red) and Gibbs sampler (blue). The magenta curves denote normal distributions centred around the ML estimates with variances as given by the empirical estimator of Duan. The true values are denoted by the black lines. (a)-(c) Estimation of the spatial polynomial coefficients $\boldsymbol{\vartheta}$. (d) Estimation of the spatially varying parameter with confidence intervals as provided by VBEM (thin dashed lines). (e) State noise precision. (f) Observation noise precision.

favourably with one another.

For this problem, the running time for the EM algorithm was 30s while that for the VBEM algorithm was 70s. The running time for the Gibbs sampler was approximately 12 hours with a total of 10,000 samples (the minimum deemed necessary for reasonable distribution estimation).[7] The computational time required by the Gibbs sampler thus exceeded that of EM and VBEM by two orders of magnitude. Since the VBEM method corroborates the results obtained by Gibbs sampling (see Figure 3.8) with such a lower requirement in terms of computational speed, the use of VBEM for a quick and accurate Bayesian identification of SPDEs is noted. The empirical estimation of parameter uncertainty using EM methods was also very quick, with Duan's method (with $l = 20$) requiring 2s.

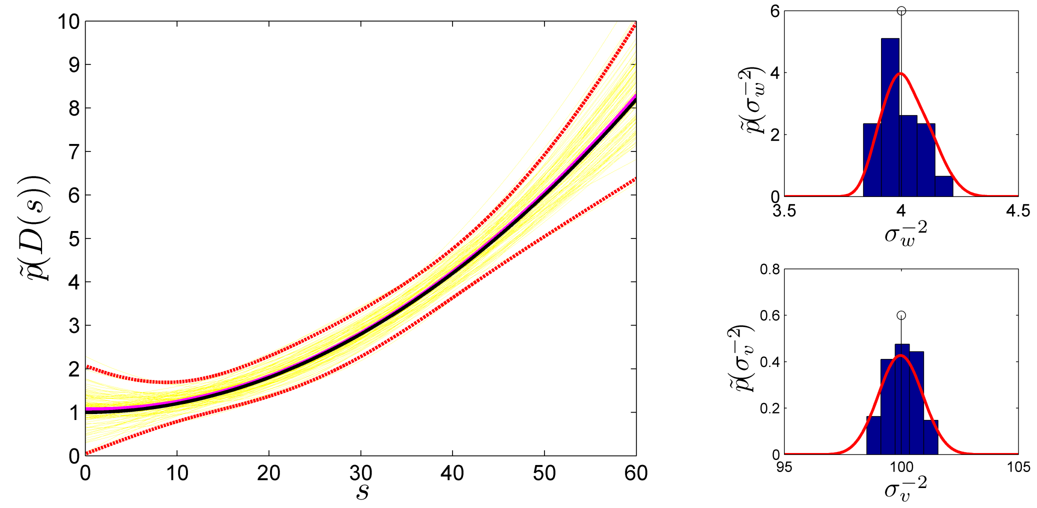
Interestingly, as shown in Table 3.1, despite being of the same order of magnitude, the variance empirical estimators of the EM algorithm and the Gibbs sampler are seen to give wider uncertainty measures than the VBEM algorithm. The VB overconfidence was somewhat expected from the comparative study in 2.3.5; a brief discussion of how this effect emerges in this case and its implication in parameter estimation of SPDEs is provided in Section 3.4.

Table 3.1: Parameter variance estimation using the EM score functions (Duan) compared with that given by VBEM and Gibbs sampling.

| | EM (Duan) | VBEM | Gibbs |
|---|---|---|---|
| var($\vartheta_1$) | 0.26 | 0.15 | 0.2 |
| var($\vartheta_2$) | 0.0023 | 0.0011 | 0.0016 |
| var($\vartheta_3$) | $7.0\times10^{-7}$ | $3.6\times10^{-7}$ | $5.4\times10^{-7}$ |

In addition to the parameter dynamics, the other two key quantities of interest which are estimated are the field and observation noise precisions in the bottom centre and bottom right panel of Figure 3.8. Here it is seen that although the mean estimates are close to the true parameter estimates, the confidence intervals are very tight around the mean estimates so that the true parameter value is treated as an outlier. These estimates are in fact biased as discussed in the next section.

[7] Simulations were carried out on an Intel®Core$^{\mathrm{TM}}$2Duo T5500 @ 1.66GHz personal computer with 2GB of RAM.



(a)

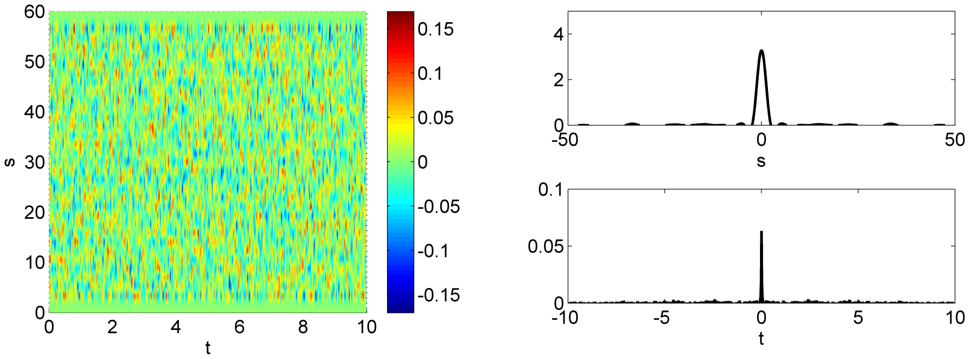
(b)

Figure 3.9: Results of 200 MC runs with simulation model basis representation equivalent to the system model basis representation. (a) True spatially varying parameter (black), mean EM estimate (magenta) and three-sigma confidence intervals as provided by VBEM (red). The profiles from the individual MC runs by EM are shown in yellow. (b) True noise precisions (black), average VB posterior (red) and histograms showing the distribution of the ML estimates.

### 3.3.4 Model mismatch analysis

Inevitably, there is a model mismatch between the reduced-order (finite dimensional) state-space model and the true (infinite dimensional) system. It is well known that model mismatch results in biased parameter estimates [180] and therefore bias will always be present when carrying out joint field-parameter estimation of SPDEs in this way. The aim of this section is to empirically show that this really is the case (and that the bias is not due to the inference methodology) and that the errors arising due to approximation are confined in some sense.

To see how the bias varies under different conditions, 200 Monte Carlo (MC) simulations were carried out using the same basis representation ($n = n^{sim} = 31, \tau = \tau^{sim} = 1.8$) and sampling interval ($\Delta_t = \Delta_t^{sim} = 0.02$) for simulation and modelling. Each MC run consisted of simulating a realisation of the SPDE using parameters (except $n, \tau$ and $\Delta_t$) fixed to those of Section 3.3.1 and applying EM and VBEM on each separate data set. A histogram of the ML estimates and the mean variational posteriors are shown in Figure 3.9 where it is graphically seen that no bias is present in both the estimated parameter dynamics and the estimated noise precisions. Of note is that the parameter



(a)

(b)

Figure 3.10: Error analysis when simulating using exact model representation. (a) Field error $z - \hat{z}$ where $\hat{z}$ is the reconstructed field using the smoothed states from the VBEM algorithm (EM algorithm gives similar results). (b) Average spatial autocorrelation across all time points (top) and average temporal autocorrelation across all spatial points (bottom).

confidence as given by the VBEM algorithm corresponds very well to that corresponding to several MC runs. Since the variance of the ML estimate is known to converge to the inverse of the Fisher information matrix (3.55), this result seems to indicate that, in this specific case, VB may be used for inferring confidence intervals with respect to the true parameter. Agreement of the Bayesian credibility intervals with frequentist confidence intervals is formalised in the Bernstein-von Mises theorem which holds for $K \to \infty$ or, informally, when the observation data set is highly informative of the parameters under study [181, 182, Chapter 4]. In the SPDE this may be violated by, for instance, increasing $\sigma_v$ as shown in Section 3.4. Field error and autocorrelation results for this problem are shown in Figure 3.10 where the desirable whiteness properties are exhibited both in space and in time. Any parameter bias and correlations in the reconstruction error introduced may thus be safely attributed to the reduction method rather than the estimation methodology.

To further study the effect of model reduction and discretisation, 100 MC runs were carried out with data generated using spatiotemporal resolutions of varying accuracy. The biases for the different parameters were then calculated. Each bias was then compared to the spread of the results of the MC runs; if the bias is more than two standard deviations of the MC trials it is considered to be significant. To find the bias on the spatially varying parameter $D(s)$, the quantity $|\hat{D}(s) - D(s)|$ was first found over a fine grid in the spatial domain and then an average evaluated. Here, $\hat{D}(s)$ denotes the mean estimated parameter over all trials.

Percentage biases together with the ratio of the biases to the standard deviations of the MC samples are given in Table 3.2.[8] It is seen that when the simulation model is exactly the same as the estimation model (with $n^{sim} = 31, \tau^{sim} = 1.8, \Delta_t^{sim} = 0.02$) the error is well within two standard deviations of the actual value for all parameters. With increasing resolution in simulation the percentage bias increases somewhat but for all cases does not exceed 10%, which may or may not be acceptable depending on the purpose of the identification exercise. The noise terms become significantly biased but, more importantly, the parameters describing the dynamics are relatively in good agreement with an error on the order of one standard deviation in all cases. In this example spatial roughness is seen to contribute significantly more to the bias than the temporal discretisation interval used in modelling. However, successive refinements of the simulation conditions do not result in an increasing bias; this is an important result required for the validity of this approach as it implies that the bias arising from the adopted dimensionality reduction method is contained in some sense.

In some cases, such as when the spatiotemporal field is decomposed with linear basis functions which construct a triangulation on $\mathcal{O}$ (a usual finite element approach), error bounds may be found for the state-space representation of Theorem 3.1, from which convergence rates may be obtained [103, 111, Section 2.5.3]. However even if bounds were found for the basis in use in this study, it is unlikely that they would be of any use in quantifying the amount of parameter bias which emerges in the inference process.

## 3.4 Parameter uncertainty measures

Like the EM algorithm, VBEM is deterministic and thus fast and memory efficient. Moreover, it readily provides a parameter uncertainty measure on the parameter dynamics $\boldsymbol{\vartheta}$. This, together with the advantage that prior information can be conveniently introduced into the estimation process, renders the use of VBEM both elegant and, arguably, a more attractive tool for joint inference. However, as can be seen from Table 3.1, the estimates given by VBEM tend to be mostly overconfident.

The reason for this lies in the missing information principle (Theorem 3.3). From (3.77) it is seen that the VB algorithm essentially attributes the posterior precision of the parameters to the expected complete-data information matrix $\mathcal{I}^c$ (3.53), which by definition from (3.46) is less than or equal to the observed-data information matrix. Hence it is $\mathcal{I}^m$ which creates the discrepancy between the two estimates, and therefore

[8] Unavailable entries in this table correspond to simulation parameters which yield instability under the explicit Euler scheme. This is a result of the stability issues concerned with the explicit Euler scheme, which are conditional on the resolution of the basis [183, Section 8.2.2].

Table 3.2: Parameter bias (in percentage) corresponding to refinement of the simulation conditions as obtained using EM. This is given together with the ratio of the bias to the standard deviation as obtained from the MC runs (> 2 corresponds to significant bias).

$D(s)$:

| $\Delta_t^{sim}$ \ $(n^{sim}, \tau^{sim})$ | (31,1.8) | (71,3.7) | (101,6.3) | (121,7.9) |
|---|---|---|---|---|
| 0.02 | 2.2% / 0.2 | 6.5% / 0.8 | na | na |
| 0.01 | 1.4% / 0.1 | 6.4% / 0.9 | na | na |
| 0.005 | 1.0% / 0.1 | 6.2% / 0.8 | 3.3% / 0.4 | 4.6% / 0.4 |
| 0.002 | 1.2% / 0.1 | 6.8% / 0.8 | 3.7% / 0.5 | 3.6% / 0.3 |

$\sigma_w^{-2}$:

| $\Delta_t^{sim}$ \ $(n^{sim}, \tau^{sim})$ | (31,1.8) | (71,3.7) | (101,6.3) | (121,7.9) |
|---|---|---|---|---|
| 0.02 | -0.2% / 0.1 | 4.3% / 1.8 | na | na |
| 0.01 | 1.6% / 0.7 | 5.9% / 2.6 | na | na |
| 0.005 | 2.4% / 1.0 | 6.5% / 2.5 | 6.3% / 2.9 | 6.8% / 2.8 |
| 0.002 | 2.6% / 1.2 | 7.0% / 3.0 | 7.0% / 3.1 | 7.6% / 3.2 |

$\sigma_v^{-2}$:

| $\Delta_t^{sim}$ \ $(n^{sim}, \tau^{sim})$ | (31,1.8) | (71,3.7) | (101,6.3) | (121,7.9) |
|---|---|---|---|---|
| 0.02 | 0.0% / 0.1 | -5.0% / 5.7 | na | na |
| 0.01 | 0.2% / 0.3 | -4.7% / 6.4 | na | na |
| 0.005 | -0.0% / 0.0 | -4.6% / 6.0 | -4.5% / 6.6 | -4.2% / 5.2 |
| 0.002 | 0.2% / 0.3 | -4.6% / -6.3 | -4.5% / 5.3 | -4.2% / 5.4 |

VBEM may be considered adequate when $\mathcal{I}^m << \mathcal{I}^c$. Interestingly, this condition is highly dependent on the quality of the observed data set $\mathcal{Y}$. To show this consider a simple LDS with state evolution equation $x_k = \vartheta x_{k-1} + w_k$ with observation equation $y_k = cx_k + v_k$, where each term takes on its usual meaning and $w_k \sim \mathcal{N}_{w_k}(0, \sigma_w^2)$. In this case, for known initial conditions and noise parameters, $\mathcal{I}^c$ is given by

$$\mathcal{I}^c(\mathcal{Y}, \vartheta) = \frac{1}{\sigma_w^2} \sum_{k=0}^{K-1} \mathbb{E}_{\tilde{p}(\mathcal{X})}[x_k^2], \tag{3.103}$$

which immediately leads to an anomaly: The increased field uncertainty may even correspond to an *increase* in the complete-data information (consider the case when noise levels generate a posterior mean which is negligible compared to the variance).

This problem (or paradox) obviously lies in missing information $\mathcal{I}^m$ which for this system is given by

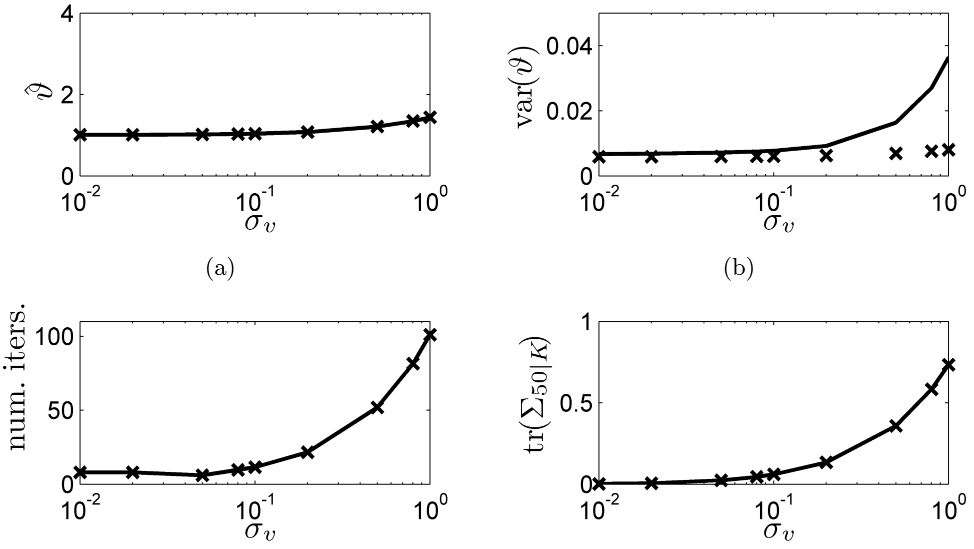
$$\mathcal{I}^m(\mathcal{Y}, \vartheta) = \frac{1}{\sigma_w^4} \mathbb{E}_{\tilde{p}(\mathcal{X})} \left[ \left( \sum_{k=1}^{K} \vartheta x_{k-1}^2 + x_k x_{k-1} \right)^2 \right]. \tag{3.104}$$

This quantity largely tends to increase with state uncertainty in the E-step due to the fourth moments involved; however this information is effectively ignored in the VBEM update (3.77). Therefore one has the result that in poorly observed systems (with states exhibiting a large variance under $\mathbb{E}_{\tilde{p}(\mathcal{X})}[\cdot]$ in relation to $\sigma_w^2$), the parameter uncertainty measures as provided by VBEM may be considered highly unrelated to the confidence of the estimate to the true parameter value.

To lay evidence to this claim tests were run on a one-dimensional spatial domain $\mathcal{O} = [0, 60]$ with $n^{sim} = 31$ basis functions, $\Delta_t^{sim} = 0.02$, $\sigma_w = 0.5$, $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$, $Qu = \int_{\mathcal{O}} k_Q(s-r)u(r)\mathrm{d}r$ with $k_Q(s) = \exp(-s^2/4)$. Only a single parameter for the unknown dynamics was considered so that $D(s) = \vartheta = 1$. It was assumed that 500 data points (in time) were gathered through the use of 101 point sensors with a sample time of $\Delta_t = 0.02$. To avoid the effects of model mismatch in this test the same model was used for both simulation and estimation. The true observation noise standard deviation $\sigma_v$ was varied from 0.01 to 1 and 200 MC runs were run for each different value of $\sigma_v$. Each of the 200 MC runs (for some $\sigma_v$) consisted of simulating a realisation of the SPDE using fixed parameters and applying EM and VBEM on each separate data set. Obviously a lower $\sigma_v$ corresponds to good-quality measurements whilst a high $\sigma_v$ corresponds to noisy measurements. Noisy sensors, in general, result in a higher estimated state variance which from (3.104) tends to increase the missing information and slow down convergence times (as a result of SEM) whilst increasing parameter uncertainty.

These effects are very evident in Figure 3.11. Indeed, VBEM (using the same priors of Section 3.3.2) is seen to maintain a fairly constant parameter uncertainty measure and does not produce the expected trend of increased parameter uncertainty with increased observation noise levels. Duan's method, on the other hand, reports a variance of nearly 5 times that given by VB for high noise levels. From Figure 3.11a it can be seen that the mean estimate drifts upwards with increased observation noise;[9] however credibility intervals given by Duan's method still enclose the true value whilst those given by VB do not. The reason for this is that in noisy conditions the complete-data information matrix is no longer a suitable approximation for the observed-data information matrix. This observation is very similar to that by Wang in [157] where it was shown that the covariance matrices obtained from VB are 'too small' when compared to those of the ML estimator on data from mixture models. The overconfidence was also noted in [184] and [133, Section 10.1.2]. Indeed, in such a scenario it is evident that the credibility intervals

[9] This is an artefact of the distribution of the ML estimate (or mode of variational posterior) which ceases to be Gaussian and becomes positively skewed, i.e., in the direction favouring a higher diffusivity. Such a phenomenon was observed also in MC runs carried out for a one-dimensional state-space model (results omitted).



(a)

(b)

(c)

(d)

Figure 3.11: (a) Average effect of increased sensor noise $\sigma_v$ over 200 MC runs on mean parameter estimate, (b) parameter uncertainty, (c) number of iterations required for convergence and (d) field uncertainty, in this case shown through the trace of the covariance matrix at the $50^{th}$ data sample. Crosses denote results obtained with VBEM. The solid lines denote results obtained with EM augmented with the variance estimator of Duan. Convergence was assumed to be reached when the change in the estimated ML (EM) or mean (VBEM) parameter was less than 1%.

given by VBEM may not be physically useful.

## 3.5 Conclusion

This chapter has applied, for the first time, several methods for the identification of linear SPDEs from continuous data. The key contributions in algorithms are:

- the rigorous formalisation of SPDEs to give finite dimensional state-space models amenable to a large array of signal processing inference tools.

- the use of frequency analysis methods for basis function placement to obtain accurate finite dimensional representations of SPDEs.

- the extension of the basic EM algorithm for SPDEs for the estimation of unknown noise parameters and provision of uncertainty measures on (if needed) all unknown parameters.

- the development of a VBEM algorithm and Gibbs sampler for SPDEs.

This chapter also provides a detailed analysis of the performance of the individual algorithms. In particular it is shown that all algorithms corroborate each other and may be of use in this scenario. It is also shown that in using any approach involving SPDE approximation and discretisation, bias will always be present in the learned parameters. Empirical studies show however that this bias is contained for a given model representation.

The use of the VB parameter posterior as an uncertainty measure is discussed. It is shown that when confidence intervals relating to the true parameter value is desired, the empirical estimator of Duan is particularly effective. It is good to note the effect of the quality of the data set on the algorithms. Sensor quality (represented by $\sigma_v$) plays a strong role in both the field and resulting parameter uncertainty and the number of iterations required for convergence by the EM based algorithms. In general a similar argument can be made for the quantity of sensors used, which affects the observability (in terms of the observability matrix) of the system. A detailed treatment of the effect of sensor quantity and quality on the proposed algorithms is noted as future work.

It is evident from the results that, based solely on computational speed and memory requirements, the EM and VBEM algorithm are the preferred methods of choice for joint field-parameter inference. Moreover, both the EM and the VBEM algorithm are seen to behave very similarly; they require the same number of iterations for convergence, give good estimates for the mean parameter value and give similar values for estimated field uncertainty (Figure 3.11). However the augmented EM algorithm is probably marginally better for simple problems such as that considered here where the true posterior distributions are not skewed; it is both slightly faster and via Duan's method is able to give uncertainty estimates which are of more practical use.

In light of the objectives set out for in Section 1.3.2, this chapter plays an important role in the framework of spatiotemporal modelling from continuous observations as it delivers a practical approach to learn a dynamic model over a continuous spatial domain from data which, as it is readily available in state-space form, can be subsequently used for control purposes. The concepts introduced in this chapter will now be extended to the rather more challenging case of when observations are events (spatiotemporal point processes).

# Chapter 4

# Field-parameter estimation from point process observations

A spatiotemporal point process is a stochastic process with samples given by countable collections of points in space and time. The random locations and times of these events are generally assumed to be generated by a (possibly) non-homogeneous[1] Poisson process governed by an intensity function $\lambda(\boldsymbol{s}, t)$ which is typically a function of a continuous secondary stochastic process $z(\boldsymbol{s}, t)$ [29]. Spatiotemporal point process models have been used to explain patterns in a wide variety of disciplines, including crop growth [35], seismology [34] and the study of disease spread in animals and humans [36].

The random intensity function of a spatiotemporal point process is commonly represented as a log-Gaussian Cox process (LGCP), i.e. $\ln \lambda(\boldsymbol{s}, t) \sim \mathcal{GP}(\cdot, \cdot)$ [185]. An advantage of this approach is the considerable ease of parameter inference through the method of contrast [36, 43]. Recently, the sigmoidal-Gaussian Cox process (SGCP) has also been introduced, which conveniently upper bounds the random intensity function and allows for MCMC techniques to be used to sample from a thinned homogeneous Poisson process [32]. However, the use of LGCPs or SGCPs to represent the intensity function over both space and time results in a model which obscures the system dynamics and which is hence not amenable to control purposes.

A different approach to point process modelling which remedies this is the dynamic systems approach, which describes the temporal evolution of the intensity as a realisation of an underlying stochastic process typically given through a state-space model [29, 31]. This approach has gained recent popularity for two reasons. First, these types

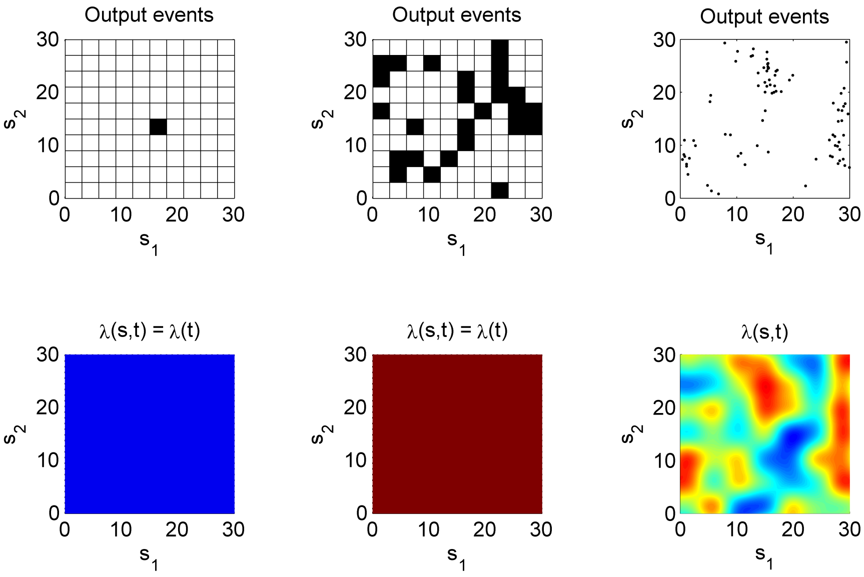
[1] In point process systems it is customary to attribute heterogeneity to temporally/spatially varying intensity. This should not be confused with the hetereogeneity as used in the previous chapters which attributed it to the spatially-varying dynamics of the spatiotemporal system.

of models can be parameterised such that appropriate inference can reveal a physical and telling interpretation to the underlying mechanism [29]; and second, the form of the state and parameter densities generally associated with this likelihood are unimodal [186], allowing a large array of approximative techniques to be realistically applied to the problem without too much compromise in accuracy. Dynamic models are particularly attractive with respect to the objectives laid out in Section 1.3.2 as they allow the approach developed in Chapter 3 to extend to the point process observations case with minor modifications.

The first development in extending the framework of Chapter 3 to point process observations is a variational Bayesian inference mechanism which can cater for the nonlinear likelihood characteristic of these systems. The inference mechanism is developed and first implemented on the simple case where a shared underlying latent state is assumed in Section 4.1. It is further shown to compare favourably against other well-established methods in Section 4.2. The second development in this chapter is the extension of the developed methodology to the general spatiotemporal case with heterogeneous intensity, described in Section 4.4. As in Chapter 3, this objective is met with the use of dimensionality reduction methods on a governing SPDE; the basis functions are now selected through frequency methods tailored to the point process case. The successful application of the proposed methodology is shown on a system governed by the stochastic heat equation in Section 4.5.

## 4.1 VBEM for homogeneous multiple channel output systems

This section is concerned with problems where the outputs from a series of independent channels are determined by an underlying intensity conditioned on a single (shared) state. Such systems are representative of spatially homogeneous processes with temporally varying intensity (Figure 4.1a and 4.1b). From the events it is required to estimate the underlying (spatially constant) intensity and infer the parameters governing the temporal evolution of this intensity. The reasons why this scenario is considered before the general heterogeneous case on a continuous-space (Figure 4.1c) is twofold: i) The variational methods developed here may find potential use in many biomedical signal processing problems, such as those concerned with neural spikes and heart beats and are hence novel in their own right. ii) The extension of the developed methodology to the usual spatially coupled spatiotemporal case follows naturally from the univariate case. Note that the simple homogeneous problem considered here may be put into a general spatiotemporal framework described later by considering a single basis function constant over the domain of interest $\mathcal{O}$.



(a) (b) (c)

Figure 4.1: Observed events (top) and underlying intensity field (bottom) over a temporal interval $\Delta_t$. Low intensity regions are marked in blue, high intensity regions in red. (a, b) Discrete-channel output systems with homogeneous intensity. (c) General spatiotemporal, heterogeneous (spatially varying state) system with observations carried out in continuous-space.

The state-space model with point process observations exhibiting a single underlying state was first proposed by Smith and Brown in [29]. This model assumes a first order autoregressive process driven by an exogenous stimulus as state dynamics and has a parameterised intensity function of an approximate Bernoulli process as its observation model. For simultaneous estimation of state and parameters of such a model, Smith and Brown derived an EM algorithm. In a recent study [186], it was shown that the expectation of the log complete-data likelihood ($\mathbb{E}_{\tilde{p}(\mathcal{X})}[\ln p(\mathcal{X},\mathcal{Y}|\boldsymbol{\theta})]$) of the state-space point process model is unimodal and highly non-Gaussian with respect to each of its parameters. The high skewness is indicative of parameter posteriors where simple ML estimates of the parameters may be quite far from the actual posterior means, motivating a Bayesian treatment of the point process model as presented later on in this section.

### 4.1.1 The observation process

Consider a set of events which are recorded on an interval $(0, T]$ from $C$ independent output channels, where each output channel corresponds to a grid location in Figure 4.1a and 4.1b. The observation length is discretised with a sampling interval $\Delta_t > 0$ so that the incoming events are represented as a sequence of binary vectors $\boldsymbol{y}_k := \boldsymbol{y}(k\Delta_t) \in \mathbb{R}^C$ where $y^c(k\Delta_t) = 1, c = 1 \ldots C$ indicates that an event has occurred at the $c^{th}$ output channel in the interval $((k-1)\Delta_t, k\Delta_t]$ and is zero otherwise. The sampling interval $\Delta_t$ is thus chosen small enough so that at most one event per sample for each output channel is present, i.e.

$$\Delta_t \in \{r; y^c(kr) \in \{0, 1\}, k \in [1, \ldots, T/r], c \in [1, \ldots, C]\}\,. \tag{4.1}$$

Given a dynamic latent state $x_k := x(k\Delta_t)$, the conditional intensity function (CIF) of the point process is defined to be of the form

$$\lambda_k^c = \lambda(k\Delta_t | x_k, \mu, \bar{\beta}^c) = \exp(\mu + \bar{\beta}^c x_k), \tag{4.2}$$

where $\mu, \bar{\beta}^c \in \mathbb{R}$. Through the conditioning on $x_k$, the CIF renders the process a non-homogeneous Poisson process [29]. The parameter $\mu$ represents a background firing rate, which for simplicity is assumed to be the same for all channels. It can be shown that the observation model (or likelihood) at the $k^{th}$ time interval in the $c^{th}$ channel is given by the approximate probability mass function defined as

$$p(y_k^c \mid x_k, \mu, \bar{\beta}^c) = [\Delta_t \lambda_k^c]^{y_k^c} \exp(-\Delta_t \lambda_k^c). \tag{4.3}$$

Equation (4.3) can be obtained from first principles by treating the binned event sequence as a series of uncorrelated Bernoulli trials [187], and is thus a realistic approximation only if (4.1) is ensured and, hence, $\Delta_t$ is sufficiently small. In practice, constraint (4.1) cannot be guaranteed before data collection, but $\Delta_t$ may be chosen such that the probability of expected arrival time within an interval $\Delta_t$ at the maximum expected intensity a priori is less than some pre-defined threshold.

Since only linear systems are considered in this thesis, the underlying state may be assumed to follow the evolution equation

$$x_k = \rho x_{k-1} + \alpha I_k + w_k, \tag{4.4}$$

where $I_k := I(k\Delta_t)$ is exogeneous and is equal to 1 if an input is present at $k\Delta_t$ and zero otherwise. $w_k := w(k\Delta_t) \in \mathbb{R}$ is additive white Gaussian noise with mean 0 and variance $\sigma_w^2 \in \mathbb{R}^+$. The initial state $x_0$ is assumed to be normally distributed with known mean $\hat{x}_{0|0}$ and variance $\sigma^2_{0|0}$. The parameters $\rho \in \mathbb{R}$ and $\alpha \in \mathbb{R}$ are the propagation constant and input gain respectively. Note that the input term is considered here since this work may be seen as a development of [29] which also considers an input term; this will be dropped when analysing the general spatiotemporal case.

Despite the simplicity (linearity) of the latent process, this model has been applied several times to represent the dynamics of a system variable, the behaviour of which is not fully understood. For instance, (4.4) has been used to successfully model the spatial receptive field of a pyramidal neuron in a rat hippocampus [30], and, more recently, to model the arousal state in subjects receiving thalamic stimulation [188].

Throughout this work it will be assumed that $\sigma_w^2$ may be fixed to a realistic value a priori but that the remaining parameters governing the firing rate $\mu$ and $\bar{\boldsymbol{\beta}} = \{\bar{\beta}^c\}_{c=1}^C$ and the state equation parameters $\alpha$ and $\rho$ are unknown. The inference problem is therefore to estimate the set of unknown parameters $\boldsymbol{\theta} \in \mathbb{R}^d, d = C + 3$ with $\boldsymbol{\theta} = \{\alpha, \rho, \mu, \bar{\beta}^1, \bar{\beta}^2, \dots, \bar{\beta}^C\}$ in addition to an underlying hidden state $x_k$ at each time point.

### 4.1.2 The VB-Laplace approach in dynamic point process systems

The variational framework for the inference in the state-space point process model is developed in a similar vein to that of Chapter 3. Let $\mathcal{X}, \mathcal{Y}$ be the set of states and observed data points up to the final time point $K = T/\Delta_t$ respectively, $\mathcal{X} = \{x_i\}_{i=0}^K$ and $\mathcal{Y} = \{\boldsymbol{y}_i\}_{i=1}^K$. As discussed in Section 2.3.3 the inference problem pivots on finding an approximation to the true posterior $p(\mathcal{X}, \boldsymbol{\theta}|\mathcal{Y}) \approx \tilde{p}(\mathcal{X}, \boldsymbol{\theta})$ such that the variational free energy is maximised. In this work, the approximate (joint) posterior is assumed to be a product of distributions

$$\tilde{p}(\mathcal{X}, \boldsymbol{\theta}) = \tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta}) = \tilde{p}(\mathcal{X})\tilde{p}(\rho, \alpha)\tilde{p}(\mu)\prod_{i=1}^{C}\tilde{p}(\bar{\beta}^i). \tag{4.5}$$

The dependency between the $\rho$ and $\alpha$ parameters may be retained since the interaction terms between them which appear when deriving the log posterior distribution are relatively easy to compute. As a result, $\alpha$ and $\rho$ are dealt with jointly and without loss of generality let $\boldsymbol{\theta} = \{(\alpha, \rho), \mu, \bar{\beta}^1, \bar{\beta}^2, \dots, \bar{\beta}^C\}$. The variational posteriors $\tilde{p}(\mathcal{X})$ and $\tilde{p}(\boldsymbol{\theta})$ are then once again given by

$$\tilde{p}(\mathcal{X}) \propto \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta})}[\ln p(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta})]), \tag{4.6}$$

$$\tilde{p}(\theta^i) \propto \exp(\mathbb{E}_{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta}^{/i})}[\ln p(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta})]), \tag{4.7}$$

where $\theta^i$ is the $i^{th}$ component in $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{/i}$ is the set of all $\boldsymbol{\theta}$ excluding $\theta^i$.

Since the VB posteriors are coupled, it is required to iterate (4.6) and (4.7) until convergence. However, because of the form of (4.3), the class of models under study in this chapter does not form part of the conjugate-exponential class of models considered in Chapter 3. As a result, the variational posteriors are, in general, not of standard form so that the required iterations are not possible analytically. The analytical computations may however be re-employed by forcing the variational posteriors to be Gaussian in form through the use of Gaussian approximation methods.

In the VB framework, the natural Gaussian approximation method is to find a Gaussian distribution such that the KL divergence between it and $\tilde{p}(\cdot)$ is a minimum. However, since the free-form variational posteriors are in generally negatively skewed, the resulting Gaussian approximation is likely to underestimate the variatonal mode which in turn may cause instability in the numerical scheme (this was found to be the case in practice). In addition, this method leads to an optimisation of $n^2$ variables in the forward and backward steps required for inferring the states. Complexity reduction mechanisms to solve this are generally only available for a special class of models [189], to which the model discussed does not belong. Since in spatiotemporal systems the state dimensionality can range into the hundreds, the dimensionality of the optimisation space may become an issue.

A somewhat simpler, but more reliable approximation method in this scenario, is the the Laplace method [133, Section 4.4] which only requires optimisation in $n$-dimensional space. Since the variational posteriors are unimodal (a proof for this proceeds on the lines of [186]) the variational-Laplace (or VB-Laplace) combination is a potential tool for approximate Bayesian inference which is both fast and computationally efficient. The VB-Laplace method may also be seen as a fixed-form VB approximation method [155] where the VB posterior densities are restricted to be Gaussian in form. A schematic overview of the VB-Laplace method adopted here is shown in Figure 4.2.

One should bear in mind that the VB-Laplace method of Figure 4.2 does *not* reduce to a simple Laplace approximation of the true posteriors $p(\mathcal{X}|\mathcal{Y})$ and $p(\boldsymbol{\theta}|\mathcal{Y})$, given here as $\bar{p}(\mathcal{X}|\mathcal{Y})$ and $\bar{p}(\boldsymbol{\theta}|\mathcal{Y})$:

$$p(\mathcal{X}, \boldsymbol{\theta}|\mathcal{Y}) \approx \tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta}) \neq \bar{p}(\mathcal{X}|\mathcal{Y})\bar{p}(\boldsymbol{\theta}|\mathcal{Y}). \tag{4.8}$$

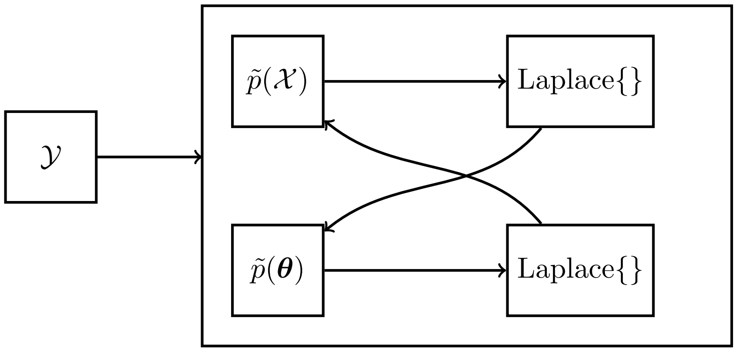


Figure 4.2: Conceptual diagram of the VB-Laplace method, where Laplace approximations are carried out on the computed VB posteriors in order to maintain recursions.

This is because the Laplace approximation is being employed on the variational posteriors which results in a joint posterior $\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta})$ with a mode which is frequently different from the mode of the true posterior [155]. This is definitely the case with distributions associated with point processes which are inherently skewed as a result of the exponentiations in the observation equation. Thus, while direct application of the Laplace method on the posteriors results in a joint Gaussian distribution centred at the mode of the true posterior distribution, the VB-Laplace scheme yields a Gaussian distribution optimally centred in relation to the majority of the probability mass.

To exemplify the distinction, assume a single event was observed at time $t = 0$ and that all the parameters are known except for $x_0$ and $\mu$. Allocate the following prior distributions

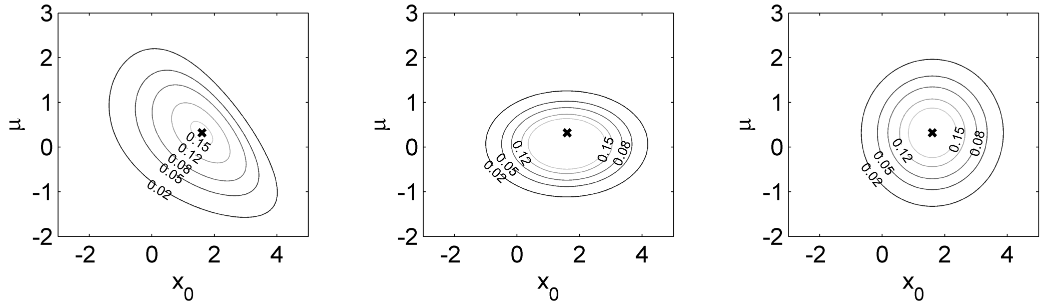
$$p(x_0) \propto \exp\left(-\frac{x_0^2}{2\sigma^2_{x_0,p}}\right) \qquad p(\mu) \propto \exp\left(-\frac{\mu^2}{2\sigma^2_{\mu,p}}\right). \tag{4.9}$$

Then from (4.3) the true posterior distribution is given by

$$p(x_0, \mu | y_0) \propto \Delta_t \exp(\mu + \bar{\beta} x_0) \exp(-\Delta_t \exp(\mu + \bar{\beta} x_0)) p(x_0) p(\mu). \tag{4.10}$$

The VB-Laplace method for this simple problem is then given by

$$\begin{aligned} \tilde{p}(x_0)^{(i+1)} &\propto \exp(\mathbb{E}_{\tilde{p}(\mu)^{(i)}}[\bar{\beta} x_0 - \Delta_t \exp(\mu + \bar{\beta} x_0)]) \exp\left(-\frac{x_0^2}{2\sigma^2_{x_0,p}}\right) \\ &\xrightarrow{Laplace} \mathcal{N}_{x_0}(\hat{x}_0^{(i+1)}, \Sigma_{x_0}^{(i+1)}), \qquad (4.11) \\ \tilde{p}(\mu)^{(i+1)} &\propto \exp(\mathbb{E}_{\tilde{p}(x_0)^{(i+1)}}[\mu - \Delta_t \exp(\mu + \bar{\beta} x_0)]) \exp\left(-\frac{\mu^2}{2\sigma^2_{\mu,p}}\right) \\ &\xrightarrow{Laplace} \mathcal{N}_{\mu}(\hat{\mu}_0^{(i+1)}, \Sigma_{\mu}^{(i+1)}). \qquad (4.12) \end{aligned}$$



(a) (b) (c)

Figure 4.3: (a) True posterior distribution (4.10). (b) VB-Laplace approximation to the true posterior distribution. (c) Minimum-risk-Laplace approximation (see Remark 4.1) to the true posterior distribution. The cross denotes the mode of the true posterior distribution. Note how the mode of the variational posterior is shifted downwards where a larger proportion of the probability mass of the true joint posterior lies.

It can then be easily shown that $\hat{x}_0^{(i+1)}$ is the solution of the nonlinear equation

$$\bar{\beta} - \bar{\beta}\Delta_t \exp(\hat{\mu}^{(i)} + \Sigma_{\mu}^{(i)}/2 + \bar{\beta}\hat{x_0}^{(i+1)}) - \hat{x_0}^{(i+1)}/\sigma^2_{x_0,p} = 0, \tag{4.13}$$

and

$$\Sigma_{x_0}^{(i+1)} = 1/(\bar{\beta}^2\Delta_t \exp(\hat{\mu}^{(i)} + \Sigma_{\mu}^{(i)}/2 + \bar{\beta}\hat{x_0}^{(i+1)}) + 1/\sigma^2_{x_0,p}). \tag{4.14}$$

Similarly one has that

$$1 - \Delta_t \exp(\hat{\mu}^{(i+1)} + \bar{\beta}(\hat{x_0}^{(i+1)} + \Sigma_{x_0}^{(i+1)}/2)) - \hat{\mu}^{(i+1)}/\sigma^2_{\mu,p} = 0, \tag{4.15}$$

and

$$\Sigma_{\mu}^{(i+1)} = 1/(\Delta_t \exp(\hat{\mu}^{(i+1)} + \bar{\beta}(\hat{x_0}^{(i+1)} + \Sigma_{x_0}^{(i+1)}/2)) + 1/\sigma^2_{\mu,p}). \tag{4.16}$$

The quantities (4.13)-(4.16) are then iterated until convergence is reached.

Carrying out the estimation with VB-Laplace under parameters $\bar{\beta} = 1$, $\Delta_t = 0.1$ and prior variances $\sigma^2_{x_0,p} = 5$ and $\sigma^2_{\mu,p} = 1$, one obtains the joint distribution $\tilde{p}(x_0)\tilde{p}(\mu)$ shown in Figure 4.3b. This is shown together with the true posterior (Figure 4.3a) and $\bar{p}(x_0)\bar{p}(\mu)$ (Figure 4.3c), where the true posterior was computed numerically. Note that the VB-Laplace method does *not* degenerate to the Laplace method on individual posterior distributions, as made clear by the difference in the positions of the modes of the approximated distributions.

**Remark 4.1** *When the true joint posterior is approximated as the product of the individual posteriors, the method is referred to as a minimum-risk approach [14, Sec-*

*tion 3.4.1]. The fundamental difference between the VB method and the minimum risk approach is the KL divergence being minimised. Whilst the VB method can be seen as a minimisation of $KL(\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta})||p(\mathcal{X},\boldsymbol{\theta}|\mathcal{Y}))$, the minimum-risk approach minimises $KL(p(\mathcal{X},\boldsymbol{\theta}|\mathcal{Y})||\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta}))$. Indeed, when attempting to evaluate KL divergence to determine which is the better approximation to the true distribution shown in Figure 4.3, the computed values favoured the method corresponding to the direction of the KL divergence employed.*

### 4.1.3 VB-Laplace for state inference

Adopting the strategy described in Section 4.1.2, state inference at the $(i+1)^{th}$ iteration, $\tilde{p}(\mathcal{X})^{(i+1)}$, may be carried out using a Laplace augmentation of the usual VB forward-backward algorithm. This filtering philosophy is similar in principle to the Laplace Gaussian filter where a Laplace approximation is taken around the mode of the true posterior distribution when the parameters are known [190]. Note that both these approaches are different from the extended Kalman filter (EKF) approach which enforces system linearity a priori.

Consider the variational forward message $\tilde{\alpha}(x_k) = \tilde{p}(x_k|\boldsymbol{y}_{1:k})$ as given in [129, Chapter 5] for a standard LDS. The idea is to approximate this quantity (at the $(i+1)^{th}$ VB iteration) as

$$\begin{aligned}\tilde{\alpha}(x_k) &\propto \int \tilde{\alpha}(x_{k-1}) \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta})^{(i)}}[\ln p(x_k|x_{k-1},\boldsymbol{\theta})p(\boldsymbol{y}_k|x_k,\boldsymbol{\theta})])\mathrm{d}x_{k-1} \\ &\xrightarrow{Laplace} \mathcal{N}_{x_k}(\hat{x}_{k|k},\sigma^2_{k|k}),\end{aligned} \tag{4.17}$$

where $p(x_k|x_{k-1},\boldsymbol{\theta}) = \mathcal{N}_{x_k}(\rho x_{k-1} + \alpha I_k, \sigma_w^2)$ and $\tilde{p}(x_{k-1}|\boldsymbol{y}_{1:k-1}) = \mathcal{N}_{x_{k-1}}(\hat{x}_{k-1|k-1},\sigma^2_{k-1|k-1})$. The product $\tilde{p}(x_{k-1}|\boldsymbol{y}_{1:k-1})\exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta})^{(i)}}[\ln p(x_k|x_{k-1},\boldsymbol{\theta})])$ is normal in $x_{k-1}$ with precision $\sigma^{*-2}_{k-1} = \sigma^{-2}_{k-1|k-1} + \mathbb{E}_{\tilde{p}(\rho)^{(i)}}[\rho^2]\sigma_w^{-2}$ and mean

$$x^*_{k-1} = \sigma^{*2}_{k-1}(\hat{x}_{k-1|k-1}\sigma^{-2}_{k-1|k-1} + \mathbb{E}_{\tilde{p}(\rho)^{(i)}}[\rho]x_k\sigma_w^{-2} - \mathbb{E}_{\tilde{p}(\rho,\alpha)^{(i)}}[\rho\alpha]I_k\sigma_w^{-2}). \tag{4.18}$$

Marginalising out $x_{k-1}$

$$\tilde{p}(x_k|\boldsymbol{y}_{1:k}) \propto \mathcal{N}_{x_k}(\tilde{x}_k,\tilde{\sigma}_k^2)\exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta})^{(i)}}[\ln p(\boldsymbol{y}_k|x_k,\boldsymbol{\theta})]), \tag{4.19}$$

where $\tilde{\sigma}_k^{-2} = \sigma_w^{-2} - \mathbb{E}_{\tilde{p}(\rho)^{(i)}}[\rho]^2\sigma^{*2}_{k-1}\sigma_w^{-4}$ and

$$\tilde{x}_k = \tilde{\sigma}_k^2\bigg(\sigma^{*2}_{k-1}\mathbb{E}_{\tilde{p}(\rho)^{(i)}}[\rho]\sigma_w^{-2}[\hat{x}_{k-1|k-1}\sigma^{-2}_{k-1|k-1} - \mathbb{E}_{\tilde{p}(\rho,\alpha)^{(i)}}[\rho\alpha]I_k\sigma_w^{-2}] + \mathbb{E}_{\tilde{p}(\alpha)^{(i)}}[\alpha]I_k\sigma_w^{-2}\bigg). \tag{4.20}$$

Hence (4.17) reduces to

$$\tilde{\alpha}(x_k) \propto \mathcal{N}_{x_k}(\tilde{x_k}, \tilde{\sigma}_k^2)\exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta})^{(i)}}[\ln p(\boldsymbol{y}_k|x_k, \boldsymbol{\theta})]) \xrightarrow{\textit{Laplace}} \mathcal{N}_{x_k}(\hat{x}_{k|k}, \sigma_{k|k}^2), \tag{4.21}$$

where, from (4.3)

$$p(\boldsymbol{y}_k|x_k, \boldsymbol{\theta}) = \prod_{c=1}^{C} \Delta_t \exp(\mu + \bar{\beta}^c x_k)^{y_k^c} \exp(-\exp(\mu + \bar{\beta}^c x_k)\Delta_t). \tag{4.22}$$

Carrying out the Laplace approximation in (4.21), the final equations governing the forward pass are obtained

$$\hat{x}_{k|k} = \tilde{x}_k + \tilde{\sigma}_k^2 \sum_{c=1}^{C} \left\{ \mathbb{E}_{\tilde{p}(\bar{\beta}^c)^{(i)}}[\bar{\beta}^c] y_k^c - \Delta_t \mathbb{E}_{\tilde{p}(\mu)^{(i)}}[\exp \mu] \frac{\mathrm{d}}{\mathrm{d}x_k} \left[ \mathbb{E}_{\tilde{p}(\bar{\beta}^c)^{(i)}}[\exp x_k \bar{\beta}^c] \right] \Big|_{x_k = \hat{x}_{k|k}} \right\}, \tag{4.23}$$

$$\sigma_{k|k}^2 = \left( \tilde{\sigma}_k^{-2} + \sum_{c=1}^{C} \left\{ \Delta_t \mathbb{E}_{\tilde{p}(\mu)^{(i)}}[\exp \mu] \frac{\mathrm{d}^2}{\mathrm{d}x_k^2} \left[ \mathbb{E}_{\tilde{p}(\bar{\beta}^c)^{(i)}}[\exp x_k \bar{\beta}^c] \right] \Big|_{x_k = \hat{x}_{k|k}} \right\} \right)^{-1}. \tag{4.24}$$

Note that the normal assumption for the variational distributions allow analytical computation of all the expectations involved in (4.18)-(4.24).

Equation (4.23) is not available in closed form and needs to be solved through a deterministic optimisation method. The prior $\hat{x}_{k|k-1}$ (obtained from the predictive distribution) can be used as a good initialisation for $\hat{x}_{k|k}$ to solve the optimisation in an efficient manner. One may even replace the state variable on the RHS by the prior to obtain a closed form solution; this rough approximation (corresponding to a single fixed-point iteration) gives adequate results with a marked decrease in computational requirements [55].

In a similar fashion, the variational backward message $\tilde{\beta}(x_k) = \tilde{p}(\boldsymbol{y}_{k+1:K}|x_k)$ is computed as

$$\tilde{\beta}(x_k) = \int \mathcal{N}_{x_{k+1}}(x'_{k+1}, \sigma'^2_{k+1}) \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta})^{(i)}}[p(x_{k+1}|x_k, \boldsymbol{\theta})]) \mathrm{d}x_{k+1}, \tag{4.25}$$

where

$$\tilde{p}(\boldsymbol{y}_{k+2:K}|x_{k+1}) \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta})^{(i)}}[\ln p(\boldsymbol{y}_{k+1}|x_{k+1}, \boldsymbol{\theta})]) \approx \mathcal{N}_{x_{k+1}}(x'_{k+1}, \sigma'^2_{k+1}), \tag{4.26}$$

and where $\tilde{p}(\boldsymbol{y}_{k+2:K}|x_{k+1}) = \mathcal{N}_{x_{k+1}}(\hat{x}_{k+1|k+2:K}, \sigma^2_{k+1|k+2:K})$. For reasons outlined later on this section, instead of the standard Laplace method as the approximation method of choice, (4.26) is approximated around the mode of the forward message ($\hat{x}_{k+1|k+1}$) rather than the mode of the true distribution to give

$$x'_{k+1} = \hat{x}_{k+1|k+1} + \sigma'^2_{k+1} \left[ \frac{\hat{x}_{k+1|k+2:K} - \hat{x}_{k+1|k+1}}{\sigma^2_{k+1|k+2:K}} + \sum_{c=1}^{C} \left\{ \mathbb{E}_{\tilde{p}(\bar{\beta}^c)^{(i)}}[\bar{\beta}^c] y^c_{k+1} - \right.\right.$$

$$\left.\left. \Delta_t \mathbb{E}_{\tilde{p}(\mu)^{(i)}}[\exp \mu] \frac{\mathrm{d}}{\mathrm{d}x_{k+1}} \left[ \mathbb{E}_{\tilde{p}(\bar{\beta}^c)^{(i)}}[\exp x_{k+1} \bar{\beta}^c] \right] \Big|_{x_{k+1} = \hat{x}_{k+1|k+1}} \right\} \right], \tag{4.27}$$

$$\sigma'^2_{k+1} = \left( \sigma^{-2}_{k+1|k+2:K} + \sum_{c=1}^{C} \left\{ \Delta_t \mathbb{E}_{\tilde{p}(\mu)^{(i)}}[\exp \mu] \frac{\mathrm{d}^2}{\mathrm{d}x^2_{k+1}} \left[ \mathbb{E}_{\tilde{p}(\bar{\beta}^c)^{(i)}}[\exp x_{k+1} \bar{\beta}^c] \right] \Big|_{x_{k+1} = \hat{x}_{k+1|k+1}} \right\} \right)^{-1}, \tag{4.28}$$

so that the mean and variance of the backward message are given by

$$\frac{\hat{x}_{k|k+1:K}}{\sigma^2_{k|k+1:K}} = \Big( \mathbb{E}_{\tilde{p}(\rho)^{(i)}}[\rho] x'_{k+1} (\sigma_w^{-2} + \sigma'^{-2}_{k+1})^{-1} \sigma_w^{-2} \sigma'^{-2}_{k+1} \tag{4.29}$$

$$+ (\sigma_w^{-2} + \sigma'^{-2}_{k+1})^{-1} \mathbb{E}_{\tilde{p}(\rho)^{(i)}}[\rho] \mathbb{E}_{\tilde{p}(\alpha)^{(i)}}[\alpha] I_{k+1} \sigma_w^{-4} - \mathbb{E}_{\tilde{p}(\rho,\alpha)^{(i)}}[\rho\alpha] I_{k+1} \sigma_w^{-2} \Big),$$

$$\sigma^2_{k|k+1:K} = (\mathbb{E}_{\tilde{p}(\rho)^{(i)}}[\rho^2] \sigma_w^{-2} - (\sigma_w^{-2} + \sigma'^{-2}_{k+1})^{-1} \mathbb{E}_{\tilde{p}(\rho)^{(i)}}[\rho]^2 \sigma_w^{-4})^{-1}. \tag{4.30}$$

The smoothed state estimate $\tilde{p}(x_k|\mathcal{Y}) \propto \tilde{\alpha}(x_k)\tilde{\beta}(x_k)$ is then Gaussian with

$$\hat{x}_{k|K} = \sigma^2_{k|K} (\hat{x}_{k|k} \sigma^{-2}_{k|k} + \hat{x}_{k|k+1:K} \sigma^{-2}_{k|k+1:K}), \qquad \sigma^2_{k|K} = (\sigma^{-2}_{k|k} + \sigma^{-2}_{k|k+1:K})^{-1}. \tag{4.31}$$

The last requirement is the computation of the cross-covariance $m_k = \mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}[x_k x_{k-1}] - \hat{x}_{k|K} \hat{x}_{k-1|K}$. This is obtained from the joint [129, Section 5.3.5]

$$\tilde{p}(x_k, x_{k-1}|\mathcal{Y}) = \tilde{\alpha}(x_{k-1}) \tilde{\beta}(x_k) \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta})^{(i)}}[p(x_k|x_{k-1}, \boldsymbol{\theta}) p(\boldsymbol{y}_k|x_k, \boldsymbol{\theta})]), \tag{4.32}$$

the natural logarithm of which (ignoring unnecessary terms) is given by

$$- \frac{(x_{k-1} - \hat{x}_{k-1|k-1})^2}{2\sigma^2_{k-1|k-1}} - \frac{(x_k - \hat{x}_{k|k+1:K})^2}{2\sigma^2_{k|k+1:K}} - \frac{\mathbb{E}_{\tilde{p}(\boldsymbol{\theta})^{(i)}}[(x_k - \rho x_{k-1} - \alpha I_k)^2]}{2\sigma_w^2}$$

$$+ \mathbb{E}_{\tilde{p}(\boldsymbol{\theta})^{(i)}} \left[ \sum_{c=1}^{C} y^c_k (\mu + \bar{\beta} x_k) - \Delta_t \exp(\mu + \bar{\beta} x_k) \right] + \dots \tag{4.33}$$

$$\approx -\frac{1}{2} \begin{pmatrix} x_{k-1} & x_k \end{pmatrix} \begin{pmatrix} a & b \\ b & d \end{pmatrix} \begin{pmatrix} x_{k-1} \\ x_k \end{pmatrix} + \dots,$$

where the approximation is carried out around the smoothed estimate. The required covariance is then given by

$$m_k = -\frac{b}{ad - b^2}, \tag{4.34}$$

where the constants $a, b$ and $d$ are found after taking second order derivatives $\partial^2/\partial x_k^2$, $\partial^2/\partial x_{k-1}^2$ and $\partial^2/\partial x_k \partial x_{k-1}$ of (4.33). These quantities are given as

$$a = \sigma_{k-1|k-1}^{-2} + \mathbb{E}_{\tilde{p}(\rho)^{(i)}}[\rho^2]\sigma_w^{-2}, \tag{4.35}$$

$$b = -\mathbb{E}_{\tilde{p}(\rho)^{(i)}}[\rho]\sigma_w^{-2}, \tag{4.36}$$

$$d = \sigma_w^{-2} + \sigma_{k|k+1:K}^{-2} + \sum_{c=1}^{C} \left\{ \Delta_t \mathbb{E}_{\tilde{p}(\mu)^{(i)}}[\exp \mu] \frac{\mathrm{d}^2}{\mathrm{d}x_k^2} \left[ \mathbb{E}_{\tilde{p}(\bar{\beta}^c)^{(i)}}[\exp x_k \bar{\beta}^c] \right] \Big|_{x_k = \hat{x}_{k|K}} \right\}. \tag{4.37}$$

**Ill-conditioning of $\tilde{\beta}(x_k)$:** A distinctive feature of the point process model with $\lambda_k^c$ as defined in (4.2) is that the mode of $\tilde{\beta}(x_k)$ is unstable in the absence of output events and inputs. For simplicity assume the parameters are known a priori and that $\bar{\beta}^c = 1, c = 1 \dots C$, then the Laplace approximation for (4.26) would yield a mode of the form

$$x'_{k+1} = \hat{x}_{k+1|k+2:K} - \gamma \exp(x'_{k+1}), \qquad \gamma > 0, \tag{4.38}$$

which always has a solution $x'_{k+1} < \hat{x}_{k+1|k+2:K}$ (since $-\gamma \exp(x'_{k+1}) < 0$). Further, it can be easily shown that under these conditions

$$\hat{x}_{k|k+1:K} = \frac{1}{\rho} x'_{k+1}, \tag{4.39}$$

so that the mode of the reverse message essentially grows exponentially negative in the absence of events and inputs when $\rho \in [0, 1)$. While $x_{k|k+1:K} \to -\infty$ is a perfectly valid solution for $\lambda_k^c$ as in (4.2) (corresponding to zero intensity), this estimate obviously creates issues when computing the smoothed estimates as a result of numerical conditioning. This anomaly, which is specific to point process systems, is particularly interesting as the unstable evolution of the backward message corresponds to that of a simple LDS when the output matrix $\boldsymbol{C} = \boldsymbol{0}$; this implies that in the absence of events in the output channels, the system is essentially *unobservable*, a concept well known in systems theory for where the states cannot be estimated from the outputs [117, Chapter 1].

Extending further the analysis to $\tilde{\alpha}(x_k)$, it immediately becomes apparent why the forward message does not follow this unstable trend. Indeed simulations showed that whilst its mode tends to go negative in the absence of inputs and events, much in the same way as in the backward pass, the mode eventually levels out. Comparing (4.23) to (4.38) one has that $\hat{x}_{k|k} < \tilde{x}_k$. However it can be easily shown from (4.20) that

$\tilde{x}_k = \rho \hat{x}_{k-1|k-1}$ and is stable, so that the forward message admits a limiting distribution with posterior mode reaching a steady state of the form

$$\hat{x}_{k|k} = \rho \hat{x}_{k|k} - \gamma \exp(\hat{x}_{k|k}), \qquad \gamma > 0, \tag{4.40}$$

giving

$$\hat{x}_{k|k} = -\gamma \frac{\exp(\hat{x}_{k|k})}{1-\rho}, \tag{4.41}$$

which has a unique solution for $\rho \in [0, 1)$. Compare this to a LDS with $\boldsymbol{C} = \boldsymbol{0}$ where the filtered state $\hat{x}_{k|k}$ simply drifts to zero but likewise is stable.[2]

One simple remedy for this problem is to ignore the backward message altogether, however this would lead to a loss in information, particularly in regions of substantial data. A better solution for the variational method in conjunction with recursions to remain viable is to approximate the probability distribution about a reasonable point. Since the forward message is guaranteed to be stable (for $\rho \in [0 \quad 1)$), its mode is considered an adequate place for effecting the approximation in (4.27) and (4.28). This has several advantages:

- If the backward message is similar to a Gaussian distribution (in the event of several events in the output channels), the approximation is guaranteed to be valid around any point, including obviously the mode of the forward message.

- In the presence of no data, biasing closer to $\hat{x}_{k|k}$ has the effect of overestimating the mode of the backward message, preventing it from going exponentially negative and thus stabilising the smoothed estimate (even though the system is effectively unobservable in these regions).

- Equations (4.27) and (4.28) yield a closed form solution. No nonlinear optimisation is required and the computational time required for the algorithm is essentially halved.

In Figure 4.4 the trajectory of the mode of $\tilde{\beta}(x_k)$ is shown when i) the Laplace approximation and ii) the Gaussian approximation about the filtered estimate is employed (for the known parameter case). The simulation was set up as per the details in Section 4.2. Note how whilst the mode tends to drift to $-\infty$ in the former approach in the absence of data, it stabilises at some negative value in the proposed approach.

**Remark 4.2** *In spatiotemporal systems the ill-conditioning may be tackled by basis func-*

[2] Corresponding analogies with the LDS can be found by working through the equations in [129, Chapter 5]

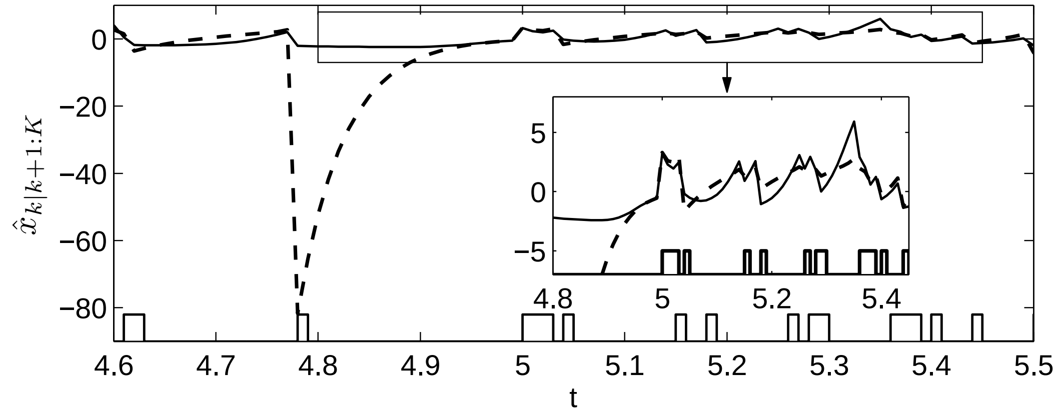


Figure 4.4: Mode of $\tilde{\beta}(x_k)$ using the Laplace approach (dashed line) and linearisations around the forward estimate (solid line). The presence or otherwise of at least one event in any output channel is indicated by the pulse train on the bottom axis. Note how the two approximation methods mostly differ in the absence of data when the system is effectively unobservable.

*tion omission in regions witnessing scarce events. These isolated events may then be captured by the background intensity constant $\mu$ which in principle may also be heterogeneous. This strategy will be adopted in Chapter 6.*

### 4.1.4 VB-Laplace for parameter inference

**Finding $\tilde{p}(\alpha)^{(i+1)}$ and $\tilde{p}(\rho)^{(i+1)}$:** Equation (4.7) gives the updates for the parameter distributions. Let the prior distributions over $\rho$ and $\alpha$ be given as $\mathcal{N}_\rho(0, \sigma^2_{\rho,p})$ and $\mathcal{N}_\alpha(0, \sigma^2_{\alpha,p})$ respectively (zero mean assumed for simplicity). As a direct consequence of the underlying linear state evolution model, the optimal variational posteriors over $\alpha$ and $\rho$ (without any need for approximation) become identical to those in a LDS, which from [129, Chapter 5] (after re-evaluation for the inclusion of $\sigma_w^{-2}$) are given by

$$\tilde{p}(\alpha)^{(i+1)} = \mathcal{N}_\alpha(\hat{\alpha}, \sigma^2_\alpha), \qquad \tilde{p}(\rho)^{(i+1)} = \mathcal{N}_\rho(\hat{\rho}, \sigma^2_\rho), \tag{4.42}$$

where

$$\hat{\alpha} = \sigma^2_\alpha \left( \frac{r_{1:K}}{\sigma^2_w} - \frac{\gamma_{1:K}\bar{\sigma}^2_\rho g_{1:K}}{\sigma^4_w} \right), \tag{4.43}$$

$$\sigma^{-2}_\alpha = \sigma^{-2}_{\alpha,p} + u_{1:K}\sigma^{-2}_w - g^2_{1:K}\bar{\sigma}^2_\rho \sigma^{-4}_w, \tag{4.44}$$

and

$$\hat{\rho} = \bar{\sigma}_\rho^2 \left( \frac{\gamma_{1:K}}{\sigma_w^2} - \frac{g_{1:K}\hat{\alpha}}{\sigma_w^2} \right), \tag{4.45}$$

$$\sigma_\rho^2 = \bar{\sigma}_\rho^2 + \frac{\bar{\sigma}_\rho^4 g_{1:K}^2 \sigma_\alpha^2}{\sigma_w^4}, \tag{4.46}$$

where the quantities containing the sufficient statistics are

$$\gamma_{k_1:k_2} = \sum_{k=k_1}^{k_2} \gamma_k = \sum_{k=k_1}^{k_2} \mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}[x_k x_{k-1}] = \sum_{k=k_1}^{k_2} [\hat{x}_{k|K}\hat{x}_{k-1|K} + m_{k|K}], \tag{4.47}$$

$$\widetilde{\lambda}_{k_1:k_2} = \sum_{k=k_1}^{k_2} \widetilde{\lambda}_k = \sum_{k=k_1}^{k_2} \mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}[x_k^2] = \sum_{k=k_1}^{k_2} [\hat{x}_{k|K}^2 + \sigma_{k|K}^2], \tag{4.48}$$

$$r_{k_1:k_2} = \sum_{k=k_1}^{k_2} r_k = \sum_{k=k_1}^{k_2} \mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}[x_k] I_k = \sum_{k=k_1}^{k_2} [\hat{x}_{k|K} I_k], \tag{4.49}$$

$$g_{k_1:k_2} = \sum_{k=k_1}^{k_2} g_k = \sum_{k=k_1}^{k_2} \mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}[x_{k-1}] I_k = \sum_{k=k_1}^{k_2} [\hat{x}_{k-1|K} I_k], \tag{4.50}$$

the sum of squared inputs is

$$u_{k_1:k_2} = \sum_{k=k_1}^{k_2} u_k = \sum_{k=k_1}^{k_2} I_k^2, \tag{4.51}$$

and

$$\bar{\sigma}_\rho^{-2} = \sigma_{\rho,p}^{-2} + \widetilde{\lambda}_{0:K-1}\sigma_w^{-2}. \tag{4.52}$$

The quantity $\mathbb{E}_{\tilde{p}(\rho,\alpha)^{(i+1)}}[\rho\alpha]$ which is required for the VB filter is given by

$$\mathbb{E}_{\tilde{p}(\rho,\alpha)^{(i+1)}}[\rho\alpha] = \bar{\sigma}_\rho^2(\gamma_{1:K}\hat{\alpha}\sigma_w^{-2} - g_{1:K}(\hat{\alpha}^2\sigma_w^{-2} + \sigma_\alpha^2\sigma_w^{-2})). \tag{4.53}$$

The mathematical details for obtaining (4.43)-(4.53) are omitted but will be shown in detail for a similar derivation in Section 5.4.1 with $\bar{\sigma}_\rho^{-2}$ replaced with $\sigma_{\rho_k|\alpha_k}^{-2}$.

**Finding $\tilde{p}(\mu)^{(i+1)}$:** The computation of the variational distribution over $\mu$ requires application of the VB-Laplace method. Ignoring terms independent of $\mu$, this posterior is given by

$$\begin{aligned}\tilde{p}(\mu)^{(i+1)} &\propto p(\mu)\exp\left(\sum_{c=1}^{C}\sum_{k=1}^{K}\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}\tilde{p}(\bar{\beta}^c)^{(i)}}[y_k^c[\mu+\bar{\beta}^c x_k]-\exp(\mu)\exp(\bar{\beta}^c x_k)\Delta_t]\right)\\ &\xrightarrow{Laplace}\mathcal{N}_\mu(\hat{\mu},\sigma_\mu^2),\end{aligned}\tag{4.54}$$

where $p(\mu)$ is the prior over $\mu$ with mean $\mu_p$ and variance $\sigma^2_{\mu,p}$. The variational posterior is restricted to be Gaussian with mean $\hat{\mu}$ and variance $\sigma^2_\mu$ by application of the standard Laplace method to obtain the expressions

$$\hat{\mu}=\mu_p+\sigma^2_{\mu,p}\sum_{k=1}^{K}\sum_{c=1}^{C}\left(y_i^c-\Delta_t\exp(\hat{\mu})\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}\tilde{p}(\bar{\beta}^c)^{(i)}}[\exp(\bar{\beta}^c x_k)]\right),\tag{4.55}$$

$$\sigma^2_\mu=\left(1/\sigma^2_{\mu_p}+\Delta_t\exp(\hat{\mu})\sum_{k=1}^{K}\sum_{c=1}^{C}\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}\tilde{p}(\bar{\beta}^c)^{(i)}}[\exp(\bar{\beta}^c x_k)]\right)^{-1}.\tag{4.56}$$

In these expressions it is required to evaluate the non-standard quantity $\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}\tilde{p}(\bar{\beta}^c)^{(i)}}[\exp(x_k\bar{\beta}^c)]$ which can be obtained as

$$\begin{aligned}\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}\tilde{p}(\bar{\beta}^c)^{(i)}}[\exp(x_k\bar{\beta}^c)] &= \int\left[\int\exp(x_k\bar{\beta}^c)\mathcal{N}_{\bar{\beta}^c}(\hat{\bar{\beta}}^c,\sigma^2_{\bar{\beta}^c})\mathrm{d}\bar{\beta}^c\right]\mathcal{N}_{x_k}(\hat{x}_{k|K},\sigma^2_{k|K})\mathrm{d}x_k\\ &= \int\exp(\hat{\bar{\beta}}^c x_k+\sigma^2_{\bar{\beta}^c}x_k^2/2)\mathcal{N}_{x_k}(\hat{x}_{k|K},\sigma^2_{k|K})\mathrm{d}x_k\\ &= \frac{1}{\sqrt{2\pi\sigma^2_{k|K}}}\int\exp(\hat{\bar{\beta}}^c x_k+\sigma^2_{\bar{\beta}^c}x_k^2/2-(x_k-\hat{x}_{k|K})^2/2\sigma^2_{k|K})\mathrm{d}x_k.\end{aligned}\tag{4.57}$$

After marginalising out $x_k$ and some algebraic manipulation the final result is obtained as

$$\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}\tilde{p}(\bar{\beta}^c)^{(i)}}[\exp(x_k\bar{\beta}^c)]=\sqrt{\frac{1}{1-\sigma^2_{\bar{\beta}^c}\sigma^2_{k|K}}}\exp\left(\frac{\hat{x}^2_{k|K}\sigma^2_{\bar{\beta}^c}+\hat{\bar{\beta}}^{c^2}\sigma^2_{k|K}+2\hat{\bar{\beta}}^c\hat{x}_{k|K}}{2(1-\sigma^2_{\bar{\beta}^c}\sigma^2_{k|K})}\right).\tag{4.58}$$

**Finding $\tilde{p}(\bar{\beta}^c)^{(i+1)}$:** Finally, the variational posterior over $\bar{\beta}^c$, ignoring terms independent of $\bar{\beta}^c$, is given by

$$\begin{aligned}\tilde{p}(\bar{\beta}^c)^{(i+1)} \propto p(\bar{\beta}^c) \exp\left(\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}\tilde{p}(\mu)^{(i+1)}}\left[\sum_{i=1}^{K} y_i^c[\mu + \bar{\beta}^c x_i] - \exp(\mu)\exp(\bar{\beta}^c x_i)\Delta_t\right]\right) \\ \xrightarrow{Laplace} \mathcal{N}_{\bar{\beta}^c}(\hat{\bar{\beta}}^c, \sigma^2_{\bar{\beta}^c}),\end{aligned} \tag{4.59}$$

where $p(\bar{\beta}^c)$ denotes the prior over $\bar{\beta}^c$ with mean $\bar{\beta}^c_p$ and variance $\sigma^2_{\bar{\beta}^c,p}$. Carrying out the required Laplace approximation gives

$$\begin{aligned}\hat{\bar{\beta}}^c &= \bar{\beta}^c_p + \sigma^2_{\bar{\beta}^c_p} \sum_{k=1}^{K} \Big( y_k^c \mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}[x_k] \\ &\qquad - \Delta_t \mathbb{E}_{\tilde{p}(\mu)^{(i+1)}}[\exp \mu] \frac{\mathrm{d}}{\mathrm{d}\bar{\beta}^c} \left[\mathbb{E}_{\tilde{p}(\mathcal{X}_K)^{(i+1)}}[\exp x_k \bar{\beta}^c]\right]\Big|_{\bar{\beta}^c = \hat{\bar{\beta}}^c} \Big), \qquad (4.60) \\ \sigma^2_{\bar{\beta}^c} &= \left( 1/\sigma^2_{\bar{\beta}^c_p} + \Delta_t \mathbb{E}_{\tilde{p}(\mu)^{(i+1)}}[\exp \mu] \sum_{k=1}^{K} \left[ \frac{\mathrm{d}^2}{\mathrm{d}\bar{\beta}^{c^2}} \mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}[\exp x_k \bar{\beta}^c] \Big|_{\bar{\beta}^c = \hat{\bar{\beta}}^c} \right] \right)^{-1}. \qquad (4.61)\end{aligned}$$

The expectations required in this case are those of log normal distributions which are easy to compute. In particular one has

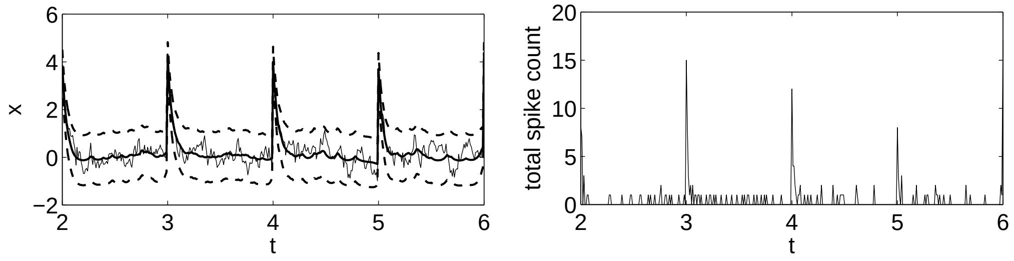
$$\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}[\exp(\bar{\beta}^c x_k)] = \exp(\bar{\beta}^c x_{k|K} + \bar{\beta}^{c^2} \sigma^2_{k|K}/2), \tag{4.62}$$

and $\mathbb{E}_{\tilde{p}(\mu)^{(i+1)}}[\exp(\mu)] = \exp(\hat{\mu} + \sigma^2_{\mu}/2)$.

## 4.2 Case study: multiple neurons driven by a shared latent process

This section discusses the performance of the above VB-Laplace algorithm when tested on synthetic data consisting of outputs of multiple neurons sharing a common hidden state. The number of neurons chosen for simulation was $C = 20$ and the response generated by a known spike input applied every 1s over a time interval of $T = 10$s with a sampling rate of 100Hz was considered (the input at time t $= 0$ was omitted). The model parameters were set to $\rho = 0.8, \alpha = 4, \mu = 0$, $\sigma^2_w = 0.05$ and $\bar{\beta}^c$ as a (uniformly) randomly generated number in the interval [0.9 1.1].

All priors on the parameters and states, except for that over $\bar{\beta}^c$, were set to normal distributions with relatively uninformative variances: $\sigma^2_{0|0} = 1$, $\sigma^2_{\rho_p} = 5, \sigma^2_{\alpha_p} = 50, \sigma^2_{\mu_p} = 1$. The prior over $\bar{\beta}^c$, on the other hand, was set to a normal distribution centred at 1 with a three-sigma confidence between 0.7 and 1.3; this was done to remedy the identifiability issues stemming from the fact that the likelihood (4.22) involves only the product $\bar{\beta}^c x_k$ (a



(a)

(b)

Figure 4.5: (a) True state (thin solid line) and mean estimated state (thick solid line) as given by the batch VB algorithm in the final iteration. The true state lies consistently within the three-sigma confidence intervals (dashed lines). (b) Total spike count from the $C$ neurons at each sampling instant.

problem related to the parameter offsetting observed in [29]). Throughout this example the nonlinear problem of finding $\hat{x}_{k|k}$, $\hat{\mu}$ and $\hat{\bar{\beta}}^c$ was carried out using the MATLAB[3] function `fzero`.

The estimation of the state variational posterior describing the latent process using the VB-Laplace algorithm can be seen in Figure 4.5 where at each time step the variational posterior's mean and three-sigma confidence limits are given. Graphical results for the corresponding estimation of the 23 unknown parameters are shown in Figure 4.6, showing rapid convergence to good estimates.

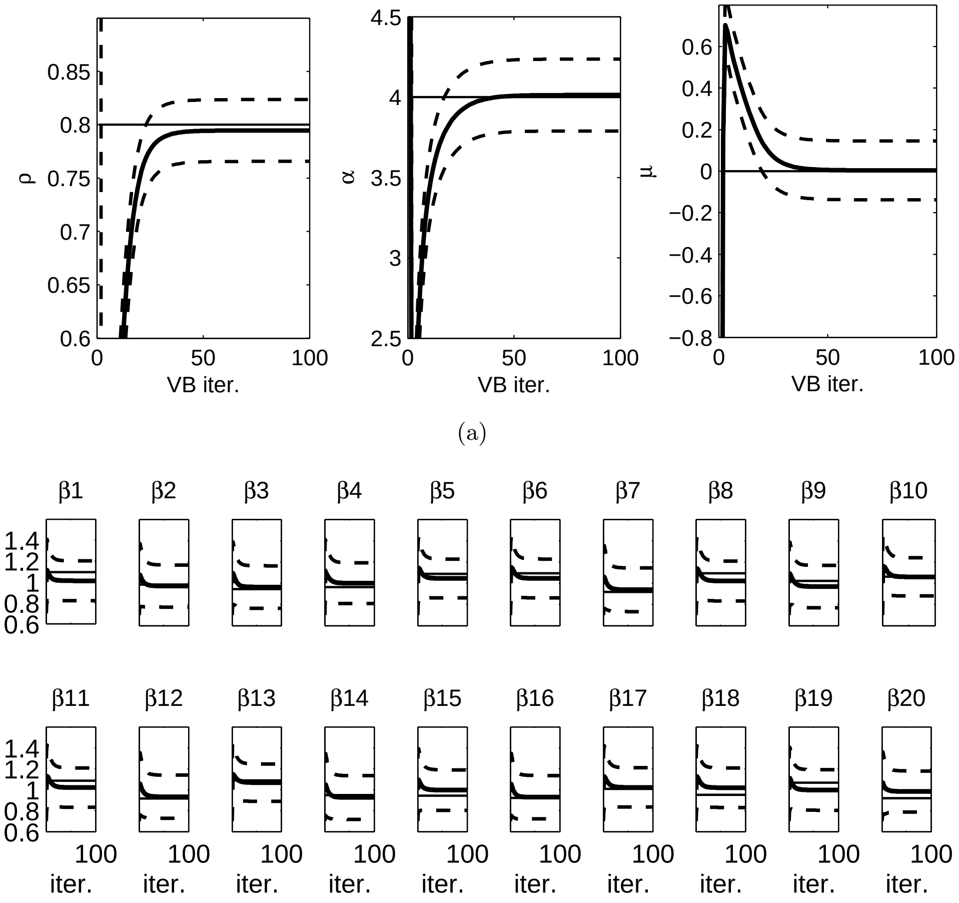
The results were further compared to those obtained using EM [29] and those given by a Gibbs sampler on the same data set.[4] To avoid identifiability issues, particularly with the EM algorithm of Smith and Brown [29] where priors are not used to prevent instability,[5] experiments were mostly carried out with $\bar{\boldsymbol{\beta}}$ fixed to its true value. Table 4.1 shows that all methods are effective in estimating parameters for this data, with the Gibbs sampler and VBEM also providing confidence intervals which are in good agreement with the true values. Note how VBEM is once again overconfident when compared to MCMC methods. It took about 6 minutes for the EM algorithm (100 iterations) and the VBEM algorithm (100 iterations), and 2 hours for the Gibbs sampler (30,000 samples) to run.[6] For the Gibbs sampler 5,000 samples were used for burn-in. Note that the discrepancy with regards to computational time between the deterministic methods and

[3] ©2011 MATLAB is a registered trademark of The MathWorks, Inc.

[4] A Gibbs sampler using the Carter and Kohn method was implemented by K. Yuan. Refer to the appendix in [55].

[5] Note that, in practice, constraining the parameter search space using MAP inference is likely to alleviate the identifiability problem associated with EM (see Section 2.3.2).

[6] Simulations carried out on an Intel® Celeron® U3600 @ 1.20GHZ with 4GB of RAM.



(a)

(b)

Figure 4.6: Mean estimates (thick solid line) and three-sigma confidence intervals (dashed lines) after 100 VBEM iterations for the parameters (a) $\rho, \alpha$, $\mu$ and (b) $\bar{\beta}^c, c = 1 \ldots 20$ using the batch VB algorithm. The parameters converge in distribution to reasonable estimates irrespective of the initial conditions and the true (thin black line) values are seen to lie well within the three-sigma confidence intervals at steady-state.

the Gibbs sampler is less than in the continuous case as a result of the required nonlinear optimisation. For this case study convergence was assessed graphically.

A more informative test of the model's performance is its ability to capture the spike train distribution. A quantitative measure of this can be achieved using the time-rescaling theorem of [191] in conjunction with a Kolmogorov-Smirnov (KS) test [28, 29]. As a goodness of fit measure, the mean squared maximum distance between the model rate and the true rate over all output channels was found. The results for this KS measure are given in Table 4.2; for completeness the results from a sliding window (SW) empirical rate-estimator of 100ms width which is often used in these applications [192] is included. The Bayesian methods (VBEM and Gibbs sampler) obtain the same goodness of fit as

Table 4.1: Parameter estimation by the EM algorithm, Gibbs sampler and VBEM algorithm. Unless stated, $\beta^c, c = 1 \dots C$ were fixed to the true values during simulation. Confidence intervals are shown at the three-sigma level. The confidence interval for $\text{avr}(\beta^c)$ is constructed on the mean variance.

| $\theta$ | True | EM | Gibbs | VBEM | VBEM (free $\bar{\boldsymbol{\beta}}$) |
|---|---|---|---|---|---|
| $\rho$ | 0.80 | 0.81 | 0.79 ± 0.06 | 0.80 ± 0.03 | 0.79 ± 0.03 |
| $\alpha$ | 4.00 | 3.91 | 3.81 ± 0.48 | 4.00 ± 0.22 | 4.01 ± 0.22 |
| $\mu$ | 0.00 | -0.01 | 0.06 ± 0.24 | -0.02 ± 0.14 | 0.00± 0.14 |
| $\text{avr}(\beta^c)$ | 1.01 | - | - | - | 1.00 ± 0.19 |

Table 4.2: Mean squared maximum KS distances for the 20 neurons with different event-rate models (lower is better) for one data set. Unless stated, $\beta^c, c = 1 \dots C$ were fixed to the true values during simulation.

| | Gibbs | VBEM | EM | VBEM (free $\bar{\boldsymbol{\beta}}$) | EM (free $\bar{\boldsymbol{\beta}}$) | SW |
|---|---|---|---|---|---|---|
| MSE | 0.0049 | 0.0049 | 0.0048 | 0.0076 | na | 0.0336 |

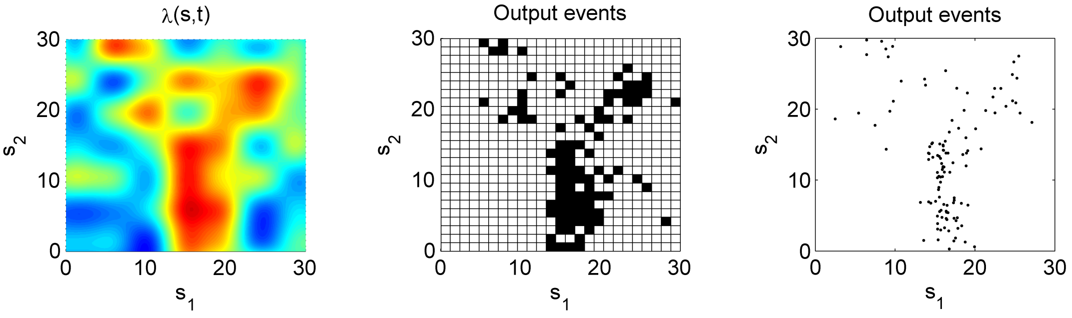
the EM algorithm and all three perform much better than the simple SW heuristic. It should be noted that in this case the EM algorithm did not converge when attempting to estimate $\boldsymbol{\beta}$ together with the other quantities.

To study whether retaining distributional information over the parameters does indeed lead to an improvement in the modelling of the spike train distribution, the VBEM and EM algorithm were run on 200 different data sets generated under the same conditions (with fixed $\boldsymbol{\beta}$). In this case the EM and VBEM algorithms were assumed to be converged when $|\hat{\rho}^{(i+1)} - \hat{\rho}^{(i)}| < 0.002$, $|\hat{\alpha}^{(i+1)} - \hat{\alpha}^{(i)}| < 0.005$ and $|\hat{\mu}^{(i+1)} - \hat{\mu}^{(i)}| < 0.004$. Since $C = 20$, a total of 4,000 output channels were analysed. For each channel the maximum distance between the model rate and the true rate using the KS tests was found both for EM and VBEM. The errors were then log-transformed so that the resulting error distributions are normal (Shapiro-Wilks test [193, Chapter 6], $p > 0.05$) with roughly equal variances (Levene's test [193, Chapter 6], $p > 0.05$). The mean log-error was -2.62 for the EM algorithm and -2.64 for the VBEM algorithm. These mean errors are remarkably close, however a 2-sample t-test showed that the reduction in error by VBEM is significant at the 10% level.

Unfortunately the t-statistic obtained has a small effect size given the large number of samples considered ($r = 0.02$) indicating that, in practice, the increased accuracy obtained by using VB-Laplace over EM for point processes may not be as advantageous as previously hoped. However, i) unlike in the continuous observation case there are no computational gains by choosing EM over VBEM as both algorithms require nonlinear optimisation (which constitutes the algorithm bottleneck). Moreover, ii) VBEM readily provides uncertainty measures over the parameters when EM does not. Finally, iii) VBEM was seen to be more stable than EM in practice, particularly when the estimation of $\boldsymbol{\beta}$ was also required, possibly as a result of the ability to constrain the parameter space through appropriate use of priors. These three advantages, together with the marginal increase in estimation accuracy, render VBEM a more suitable tool for point process systems than the conventional EM algorithm.

## 4.3 Framework for heterogeneous spatiotemporal point processes

In order to facilitate the extension of the methodology from homogeneous discrete-space output systems to heterogeneous continuous-space output systems it is worth giving an overview of a framework for heterogeneous discrete output systems. See Figure 4.7 for a qualitative comparison between the latter two representations.



(a) (b) (c)

Figure 4.7: (a) Underlying heterogeneous intensity field over a temporal interval $\Delta_t$ with observed events under (b) a discrete-space representation and (c) a continuous-space representation.

### 4.3.1 Discrete-space representations

Unlike Section 4.1, where the ordering (i.e. the actual physical position) of each $y_k^c$ was irrelevant as a result of system homogeneity, in heterogeneous discrete output systems each $y_k^c$ denotes an actual, fixed, physical location (usually taken to be a grid area) in space, say $\mathcal{O}^c$. Let $z(t) \in H$ denote the underlying evolving field and $z(t) = z(k\Delta_t) := z_k$.

Then the intensity $\lambda_k^c$ generating the observations at $y_k^c$ is conditional on the average of $z_k$ in $\mathcal{O}^c$; specifically

$$\lambda_k^c = \exp(\mu + \bar{\beta} z_k^c), \tag{4.63}$$

where

$$z_k^c = \int_{\mathcal{O}^c} z_k \mathrm{d}\mathbf{s}, \tag{4.64}$$

and where it is assumed that $\bar{\beta}^c = \bar{\beta}, c = 1 \dots C$ (spatially constant). Now, if $z_k^c$ is decomposed using a set of basis functions as in (3.3) one has

$$z_k^c = \int_{\mathcal{O}^c} \sum_{i=1}^{n} x_{i,k} \phi_i \mathrm{d}\mathbf{s} = \boldsymbol{x}_k^T \bar{\boldsymbol{\phi}}_c, \tag{4.65}$$

where

$$\bar{\boldsymbol{\phi}}_c = \left[ \int_{\mathcal{O}^c} \phi_1(\mathbf{s}) \mathrm{d}\mathbf{s} \quad \int_{\mathcal{O}^c} \phi_2(\mathbf{s}) \mathrm{d}\mathbf{s} \quad \dots \quad \int_{\mathcal{O}^c} \phi_n(\mathbf{s}) \mathrm{d}\mathbf{s} \right]^T. \tag{4.66}$$

The likelihood given in (4.22) is then represented as

$$p(\boldsymbol{y}_k | x_k, \boldsymbol{\theta}) = \prod_{c=1}^{C} \Delta_t \exp(\mu + \bar{\beta} \boldsymbol{x}_k^T \bar{\boldsymbol{\phi}}_c)^{y_k^c} \exp(-\exp(\mu + \bar{\beta} \boldsymbol{x}_k^T \bar{\boldsymbol{\phi}}_c) \Delta_t). \tag{4.67}$$

Note that since the discretisation is usually carried out a priori, $\{\mathcal{O}^c\}_{c=1}^C$ is known and hence $\{\bar{\boldsymbol{\phi}}_c\}_{c=1}^C$ is also known a priori. Equation (4.67) is a very powerful representation; it implies that if $z(\boldsymbol{s}, t)$ is a dynamic spatiotemporal field (obeying for instance an SPDE or an SIDE) which can be projected into $H^n$ and discretised in time using a finite difference scheme with interval $\Delta_t$ with reasonable accuracy, then both the field and the underlying spatiotemporal parameters may be estimated by simply extending the methodology of Section 4.1 to the multivariate case.

However this representation has limited applicability as it is highly restricted by the fact that a mesh needs to be defined a priori and by constraint (4.1) recall that only one event is allowed per output channel at any given instant. Attempts to implement this methodology are likely to result in a fine mesh which is too computationally and memory intensive for the developed methods, so that discretisations which sometimes violate (4.1) are used, thus contributing to a loss in information (compare Figure 4.7b with Figure 4.7c for an example of this effect). The ignoring or 'glossing over' of hot spots in intensity estimation obviously also has its consequences in parameter estimation. These problems are remedied by treating the observations in continuous-space. Today, with the advent of remote-sensing technologies and the global positioning system (GPS), where the precise coordinates of the output event may be known, a framework handling the exact spatial locations is bound to be of more considerable use in the application domain.

### 4.3.2 Continuous-space representation

To extend the framework to continuous-space scenarios, it is required to derive the likelihood of a temporally sampled spatiotemporal point process. This is given in the following theorem.

**Theorem 4.1** *Let $\mathcal{P}$ denote a collection of points in space and time over a domain $\mathcal{O}\times\mathcal{T}$ with $\mathcal{O} \subset \mathbb{R}$,$\mathcal{T} \subset \mathbb{R}^+$ and let $\mathcal{P}_k$ denote a spatial process consisting of points in the time interval $((k-1)\Delta_t, k\Delta_t]$. The likelihood of $\mathcal{P}_k$ is then given as*

$$p(\mathcal{P}_k|\lambda_k(s)) = \prod_{s_j\in\mathcal{P}_k} \Delta_t\lambda_k(s_j)\exp\left(-\Delta_t\int_{\mathcal{O}}\lambda_k(s)\,ds\right). \tag{4.68}$$

*Proof.* Consider a discrete approximation of the spatiotemporal domain $\mathcal{O}\times\mathcal{T}$ with $K$ frames and $J$ grid locations in each frame, such that there is no more than one event in every space-time compartment. Let each compartment be of area $\Delta = \Delta_t\Delta_s$ and let $c_{j,k}$ denote the compartment at the $k^{th}$ time point and $j^{th}$ spatial location.[7] Then, following the definition of a Poisson distribution, for any $c_{j,k}$ [194, Chapter 2]

$$Pr(\text{event in } c_{j,k}) = \bar{\lambda}(s_j,t_k)\exp(-\bar{\lambda}(s_j,t_k)), \tag{4.69}$$
$$Pr(\text{no event in } c_{j,k}) = \exp(-\bar{\lambda}(s_j,t_k)), \tag{4.70}$$

where $t_k = k\Delta_t, s_j = j\Delta_s$ and

$$\bar{\lambda}(s_j,t_k) = \int\int_{c_{j,k}} \lambda(s,t)\mathrm{d}t\mathrm{d}s. \tag{4.71}$$

The probability of obtaining points in compartments with space-time indices in the set $\mathcal{W}$ and no points elsewhere is then given by

$$\prod_{j,k\in\mathcal{W}} \bar{\lambda}(s_j,t_k)\exp(-\bar{\lambda}(s_j,t_k)) \prod_{j,k\in\mathcal{W}^c} \exp(-\bar{\lambda}(s_j,t_k)) \tag{4.72}$$
$$= \prod_{j,k\in\mathcal{W}} \bar{\lambda}(s_j,t_k)\exp\left(-\int_{\mathcal{O}\times\mathcal{T}}\lambda(s,t)\mathrm{d}s\mathrm{d}t\right). \tag{4.73}$$

Since $\mathcal{P}$ is a Poisson process, it follows that any subset of $\mathcal{P}$ defined on a finite interval in $\mathcal{O}\times\mathcal{T}$ is also a Poisson process. In particular let $\mathcal{P}_k \subset \mathcal{P}$ be defined on $(t_{k-1}, t_k]\times\mathcal{O}$

[7] One spatial dimension is treated without any loss of generality.

and assume that $\lambda(s,t)$ is constant over a small interval $\Delta_t$. Then

$$Pr(\mathcal{P}_k|\lambda_k(s)) = \prod_{j\in\mathcal{W}_k} \Delta_t \bar{\lambda}_k(s_j) \exp\left(-\Delta_t \int_{\mathcal{O}} \lambda_k(s)\mathrm{d}s\right), \tag{4.74}$$

where $\mathcal{W}_k$ is set of indices of compartments in $(t_{k-1}, t_k]$ containing an event and $\bar{\lambda}_k(s_j) = \int_{s_{j-1}}^{s_j} \lambda_k(s)\mathrm{d}s$. Now $\bar{\lambda}_k(s_j) \approx \Delta_s \lambda_k(s_j)$ so that

$$Pr(\mathcal{P}_k|\lambda_k(s)) = \Delta_s^{|\mathcal{W}_k|} \prod_{j\in\mathcal{W}_k} \Delta_t \lambda_k(s_j) \exp\left(-\Delta_t \int_{\mathcal{O}} \lambda_k(s)\mathrm{d}s\right), \tag{4.75}$$

where in this case $|\cdot|$ denotes set cardinality. Equation (4.75) gives the probability of observing the events $\mathcal{P}_k$ on finite subsets of spatial size $\Delta_s$. To obtain the probability density of the observed events, as in the temporal case [194, Chapter 2] divide by $\Delta_s^{|\mathcal{W}_k|}$ to give (4.68). ■

It is intuitive to compare (4.68) with the discrete-space likelihood

$$\prod_{c=1}^{C} [\Delta_t \lambda_k^c]^{y_k^c} \exp(-\Delta_t \lambda_k^c). \tag{4.76}$$

The first part of the likelihoods is given by $\prod_{c=1}^{C}[\Delta_t\lambda_k^c]^{y_k^c}$ in discrete-space and $\prod_{\boldsymbol{s}_j\in\mathcal{P}_k} \Delta_t\lambda_k(\boldsymbol{s}_j)$ in continuous-space. Whilst the former allows at most one event per compartment, the latter permits numerous events at a close proximity to each other. The problem of mesh selection a priori is therefore omitted.

Of even more importance is the difference in the second part of the likelihoods, $\prod_{c=1}^{C} \exp(-\Delta_t\lambda_k^c)$ as compared with $\exp(-\Delta_t \int_{\mathcal{O}} \lambda_k(\boldsymbol{s}))$. Although numerical methods are still required to compute the integral, the spatial discretisation employed only needs to be sufficiently fine so that numerical integration of fields represented by $\boldsymbol{\phi}$ are adequately computed, i.e. given a field's frequency content (to be made precise in Section 4.5.2) and an appropriate set of basis functions, a numerical integration scheme must be defined which approximates $\int_{\mathcal{O}} \lambda_k(\boldsymbol{s})\mathrm{d}\boldsymbol{s}$ accurately. The numerical scheme is thus independent of the number of points, or even the proximity of the points to each other (the desire of such a property is very evident in the discussion of Illian and Simpson in [20]).

This feature of the continuous-space representation is of vital importance in highly heterogeneous data sets. For instance, the case considered in Chapter 6 contains both regions where events are separated by hundreds of kilometers at every time point, and regions where events are only separated by a few kilometers. With the discrete-space representation a very fine mesh would have had to be used in order to safeguard (4.1) over the entire spatial domain and would have required a grid several orders of magnitude finer than that necessary for accurate integration in (4.68). The modelling approach would have not been viable for inference in this case.

## 4.4 VB-Laplace for SPDEs from point process observations

Following the strategy described for SPDEs with continuous observations in Chapter 3, this section now proceeds to discuss basis function placement and inference in the context of point process observations by building on the work carried out in Section 4.1.

### 4.4.1 Basis function placement from point processes

Consider a CIF of the form

$$\lambda_k = \exp(\mu + z_k), \tag{4.77}$$

which implies, without loss of generality, that $\bar{\beta} = 1$. As shown in Section 3.3.2, the basis functions should be able to reconstruct the field with sufficient accuracy, and this is assured by choosing basis functions with a bandwidth greater than that of the field as in (3.25). When considering linear time series observations, the frequency content of the field is easily found by taking the Fourier transform of the sensor readings. But how can one determine the frequency content of a field when only events in space are observed?

To answer this consider the following two scenarios; if throughout a time interval $\Delta_t$ points are placed randomly in space (homogeneously distributed), they are said to be spatially uncorrelated and the intensity field may be assumed to be constant (as is seen in Figure 4.1a and 4.1b). If on the other hand the points tend to cluster together (as in Figure 4.1c), then the points are correlated, and the type of clusters (very dense hot spots, or clusters in which the events are quite sparse) determine the range of spatial interactions. The type of correlation is then, in turn, representative of the underlying field's frequency content by the *autocorrelation theorem* which states the following:

**Theorem 4.2 (Autocorrelation theorem [195, Chapter 6])** *If $Z(\nu)$ is the Fourier transform of $z(s)$ then its autocorrelation function $\psi(s)$ has a Fourier transform $|Z(\nu)|^2$, the power spectral density (PSD) of the signal.*

In a more precise sense, the correlation between points in space is studied using a quantity commonly referred to in point processes as the *pair autocorrelation function (PACF)* which is defined through the first and second moment measures of the random intensity.

**Definition 4.1** *Let* $\lambda_k^{(1)}(\boldsymbol{s}) = \mathbb{E}[\lambda_k(\boldsymbol{s})]$ *and* $\lambda_{k,k}^{(2)}(\boldsymbol{s},\boldsymbol{r}) = \mathbb{E}[\lambda_k(\boldsymbol{s})\lambda_k(\boldsymbol{r})]$*, then*

$$g_{k,k}(\boldsymbol{s},\boldsymbol{r}) = \frac{\lambda_{k,k}^{(2)}(\boldsymbol{s},\boldsymbol{r})}{\lambda_k^{(1)}(\boldsymbol{s})\lambda_k^{(1)}(\boldsymbol{r})}, \tag{4.78}$$

*is known as the PACF.*

The quantity $g_{k,k}(\boldsymbol{s},\boldsymbol{r})$ is particularly useful because of its relationship to the statistics of log Gaussian random fields.

**Definition 4.2** *Let* $z_k(\boldsymbol{s}) \sim \mathcal{GP}(\hat{z}_k(\boldsymbol{s}), \sigma_k^2\psi_k(\boldsymbol{s},\boldsymbol{r}))$*, where* $\psi$ *is a correlation function with* $\psi(\boldsymbol{s},\boldsymbol{s}) = 1$*, then* $\lambda_k(\boldsymbol{s}) = \exp(\mu + z_k(\boldsymbol{s}))$ *is termed a log Gaussian Cox process.*

LGCPs [185] are typically spatial-only fields used to describe the log normal properties of the random intensity function. It follows that since $z_k(\boldsymbol{s})$ is reconstructed from a linear Gaussian evolution equation in this thesis (Theorem 3.1), each $z_k(\boldsymbol{s})$ is a GP and each $\lambda_k(\boldsymbol{s})$ is a LGCP.

**Lemma 4.1** *The PACF of a LGCP is related to the covariance function of the GP through*

$$g_{k,k}(\boldsymbol{s},\boldsymbol{r}) = \exp(\sigma_k^2\psi_k(\boldsymbol{s},\boldsymbol{r})) \tag{4.79}$$

*Proof.* Consider the form of $\lambda_k$ in (4.77). Since $\lambda_k(\boldsymbol{s})$ is log normal[8] one has that

$$\lambda_k^{(1)}(\boldsymbol{s}) = \mathbb{E}[\lambda_k(\boldsymbol{s})] = \exp(\mu + \hat{z}_k(\boldsymbol{s}) + \sigma_k^2/2). \tag{4.80}$$

Further

$$\begin{aligned}\lambda_{k,k}^{(2)}(\boldsymbol{s},\boldsymbol{r}) &= \mathbb{E}[\exp(\mu + z_k(\boldsymbol{s}) + \mu + z_k(\boldsymbol{r}))] \\ &= \exp(2\mu)\mathbb{E}[\exp(z_k(\boldsymbol{s}) + z_k(\boldsymbol{r}))]. \end{aligned}\tag{4.81}$$

Now the first and second moments of the exponent are given as

$$\mathbb{E}[z_k(\boldsymbol{s}) + z_k(\boldsymbol{r})] = \hat{z}_k(\boldsymbol{s}) + \hat{z}(\boldsymbol{r}), \tag{4.82}$$

$$\begin{aligned}\mathbb{E}[(z_k(\boldsymbol{s}) + z_k(\boldsymbol{r}))^2] &= \mathbb{E}[z_k(\boldsymbol{s})^2 + z_k(\boldsymbol{r})^2 + 2z_k(\boldsymbol{s})z_k(\boldsymbol{r})] \\ &= \hat{z}_k(\boldsymbol{s})^2 + \hat{z}_k(\boldsymbol{r})^2 + 2\hat{z}(\boldsymbol{s})\hat{z}(\boldsymbol{r}) + 2\sigma_k^2 + 2\sigma_k^2\psi_k(\boldsymbol{s},\boldsymbol{r}), \end{aligned}\tag{4.83}$$

[8] Recall that if $x$ is normally distributed with mean $\mu$ and variance $\sigma^2$, then $\mathbb{E}[\exp(x)] = \exp(\mu+\sigma^2/2)$.

so that $\text{var}[z_k(\boldsymbol{s}) + z_k(\boldsymbol{r})] = 2\sigma_k^2 + 2\sigma_k^2\psi_k(\boldsymbol{s},\boldsymbol{r})$. Therefore

$$\lambda_{k,k}^{(2)}(\boldsymbol{s},\boldsymbol{r}) = \exp(2\mu + \hat{z}_k(\boldsymbol{s}) + \hat{z}(\boldsymbol{r}) + \sigma_k^2 + \sigma_k^2\psi_k(\boldsymbol{s},\boldsymbol{r})), \tag{4.84}$$

$$\lambda_k^{(1)}(\boldsymbol{s})\lambda_k^{(1)}(\boldsymbol{r}) = \exp(2\mu + \hat{z}_k(\boldsymbol{s}) + \hat{z}(\boldsymbol{r}) + \sigma_k^2). \tag{4.85}$$

Substituting (4.84) and (4.85) in (4.78) gives the required result. ■

For simplicity in this work it will be assumed that the correlation function is translation invariant, i.e. that $\psi(\boldsymbol{s},\boldsymbol{r}) = \psi(\boldsymbol{s}-\boldsymbol{r})$, and moreover that it is isotropic, i.e. that $\psi(\boldsymbol{s},\boldsymbol{r}) = \psi(||\boldsymbol{s}-\boldsymbol{r}||) = \psi(\upsilon)$. It thus follows that the PACF is also translation invariant and istropic and that, from Lemma 4.1,

$$\ln g_{k,k}(\upsilon) \propto \psi_k(\upsilon). \tag{4.86}$$

Recall that since $\psi(\upsilon)$ is an autocorrelation function, it bears information on the frequency content of the spatial field. The problem is therefore to find $g_{k,k}(\upsilon)$, or rather the average over all time points $\bar{g}_{k,k}(\upsilon)$, from the data with the use of non-parametric analysis (use of the average is possible under the assumption of temporal stationarity). Fortunately this is today standard in the point process literature and the reader is referred to Appendix C for details. Once $\bar{g}_{k,k}(\upsilon)$ is found, Fourier analysis may then be used to find $\nu_c$, after which the approach for basis function placement of Section 3.1.3 may be readily implemented.

**Remark 4.3** *A suitable form for $\lambda_k^{(1)}(\boldsymbol{s})$ needs to be constructed for the method to work. In the ensuing example and in Chapter 6 first-order spatial stationarity will be assumed so that*

$$\lambda_k^{(1)} = \frac{N_k}{|\mathcal{O}|} \in \mathbb{R}^+, \tag{4.87}$$

*where $N_k$ is the number of points at the $k^{th}$ time frame and $|\mathcal{O}|$ is the area of the space under consideration. In some cases, linear or polynomial functions for $\lambda_k^{(1)}(\boldsymbol{s})$ may be a more suitable choice [35, 42, Section 5.6].*

### 4.4.2 Inference

If $z(t)$ follows an SPDE of the form in (2.20) it can be approximated using the discrete-time finite dimensional representation of SPDEs in Theorem 3.1 and may hence be represented as a linear equation of the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}(\boldsymbol{\vartheta})\boldsymbol{x}_k + \boldsymbol{w}_k. \tag{4.88}$$

Now assume the observations are isolated events governed by an underlying CIF of the form

$$\lambda_k = \exp(\mu + \bar{\beta} z_k). \tag{4.89}$$

Under the Galerkin approximation, the likelihood of (4.68) is given as

$$p(\{\boldsymbol{y}_k\}|\boldsymbol{x}_k, \mu, \bar{\beta}) = \prod_{\boldsymbol{s}_j \in \mathcal{W}_k} \Delta_t \exp(\mu + \bar{\beta}\boldsymbol{x}_k^T \boldsymbol{\phi}(\boldsymbol{s}_j)) \exp\left(-\Delta_t \int_{\mathcal{O}} \exp(\mu + \bar{\beta}\boldsymbol{x}_k^T \boldsymbol{\phi}(\boldsymbol{s}))\mathrm{d}\boldsymbol{s}\right), \tag{4.90}$$

where $\{\boldsymbol{y}_k\} = \mathcal{P}_k$ denotes the (time-varying size) set of coordinates in the spatial point process and $\boldsymbol{\phi}(\boldsymbol{s}_j)$ is the vector of basis functions at event location $\boldsymbol{s}_j$. The set of equations (4.88) and (4.90) describe a linear system with point process observations having unknown states $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=0}^K$, $\boldsymbol{x}_0 \sim \mathcal{N}_{\boldsymbol{x}_0}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, observed data $\mathcal{Y} = \{\{\boldsymbol{y}_i\}_{i=1}^K\}$ and unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\vartheta}, \mu, \sigma_w^{-2}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0^{-1})$. In order to simplify the exposition, no inputs are considered in the model. Further the parameters $\mu$, $\bar{\beta}$, $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0^{-1}$ and $\sigma_w^{-2}$ are assumed to be known a priori; however similar arguments as shown in Chapter 3 may be employed to estimate these quantities too. The problem of SPDE identification from point processes may thus be solved by estimating the states governing the field statistics $\mathcal{X}$ and the unknown quantities $\boldsymbol{\vartheta}$ given the observed data.

Approximate the (joint) posterior in the usual way

$$\tilde{p}(\mathcal{X}, \boldsymbol{\theta}) = \tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta}) = \tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\vartheta}). \tag{4.91}$$

The optimal choice for the variational posteriors $\tilde{p}(\mathcal{X})$ and $\tilde{p}(\boldsymbol{\vartheta})$ is then given by

$$\tilde{p}(\mathcal{X}) \propto \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})}[\ln p(\mathcal{X}, \mathcal{Y}, \boldsymbol{\vartheta})]), \tag{4.92}$$

$$\tilde{p}(\boldsymbol{\vartheta}) \propto \exp(\mathbb{E}_{\tilde{p}(\mathcal{X})}[\ln p(\mathcal{X}, \mathcal{Y}, \boldsymbol{\vartheta})]). \tag{4.93}$$

**Finding $\tilde{p}(\mathcal{X})^{(i+1)}$:** In a similar vein as (4.17) one has

$$\begin{aligned}\tilde{\alpha}(\boldsymbol{x}_k) &\propto \int \tilde{\alpha}(\boldsymbol{x}_{k-1}) \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})^{(i)}}[\ln p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{\vartheta}) p(\{\boldsymbol{y}_k\}|\boldsymbol{x}_k)]) \mathrm{d}\boldsymbol{x}_{k-1} \\ &\xrightarrow{Laplace} \mathcal{N}_{\boldsymbol{x}_k}(\hat{\boldsymbol{x}}_{k|k}, \boldsymbol{\Sigma}_{k|k}),\end{aligned} \tag{4.94}$$

which, following the derivation for the univariate case,[9] results in

$$\tilde{\alpha}(\boldsymbol{x}_k) \propto \mathcal{N}_{\boldsymbol{x}_k}(\tilde{\boldsymbol{x}}_k, \tilde{\boldsymbol{\Sigma}}_k) \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})^{(i)}}[\ln p(\{\boldsymbol{y}_k\}|\boldsymbol{x}_k, \boldsymbol{\vartheta})]) \xrightarrow{Laplace} \mathcal{N}_{\hat{\boldsymbol{x}}_k}(\hat{\boldsymbol{x}}_{k|k}, \boldsymbol{\Sigma}_{k|k}). \tag{4.95}$$

[9] Without redefining, uppercase and bold symbols are used to represent the same quantities which are now matrices and vectors respectively.

The approximation requires evaluation of the gradient (recall $\mu$ and $\bar{\beta}$ are assumed to be known) of the logarithm of (4.90) for optimisation. This is given by

$$\frac{\partial}{\partial \boldsymbol{x}_k} \ln p(\{\boldsymbol{y}_k\}|\boldsymbol{x}_k) = \sum_{\boldsymbol{s}_j \in \mathcal{P}_k} \bar{\beta}\boldsymbol{\phi}(\boldsymbol{s}_j) - \Delta_t \exp(\mu)\bar{\beta} \int_{\mathcal{O}} \boldsymbol{\phi}(\boldsymbol{s}) \exp(\bar{\beta}\boldsymbol{\phi}(\boldsymbol{s})^T \boldsymbol{x}_k) \mathrm{d}\boldsymbol{s}. \tag{4.96}$$

Scaled conjugate gradient (SCG) can be effectively employed to find the local maxima in $n$-dimensional space. To accelerate the search, the SCG is initialised using the predictive distribution from the Kalman filter assuming point estimate parameters, $\hat{\boldsymbol{x}}_{k|k-1}$, which in practice is a good estimate. Further, the Hessian used to obtain the variance is given by

$$\frac{\partial^2}{\partial \boldsymbol{x}_k \partial \boldsymbol{x}_k^T} \ln p(\{\boldsymbol{y}_k\}|\boldsymbol{x}_k) = -\Delta_t \exp(\mu)\bar{\beta}^2 \int_{\mathcal{O}} \boldsymbol{\phi}(\boldsymbol{s})\boldsymbol{\phi}(\boldsymbol{s})^T \exp(\bar{\beta}\boldsymbol{\phi}(\boldsymbol{s})^T \boldsymbol{x}_k) \mathrm{d}\boldsymbol{s}. \tag{4.97}$$

Numerical integration methods are required to compute the integrals in (4.96) and (4.97) but the order of the integral is limited to the dimensionality of the physical space under consideration which is never more than 3 and, as alluded to earlier, the computational time required to evaluate the integral is determined by the choice of basis functions and not by the locations of the individual points.

The gradient and the Hessian are also required in the computation of the backward message $\tilde{\beta}(\boldsymbol{x}_k)$, which in turn is combined with $\tilde{\alpha}(\boldsymbol{x}_k)$ to produce the smoothed estimate as in (4.31). A small note is required on the computation of the cross covariance matrix $\boldsymbol{M}_k$; since $a, b$ and $d$ in (4.33) become block matrices in the multivariate cases Schur complements [196] are required for its evaluation [129, Chapter 5]. The full mathematical details for the variational Kalman smoother for the SPDE with point process observations are given in Algorithm A.3.

**Finding $\tilde{p}(\boldsymbol{\vartheta})^{(i+1)}$:** Since $\boldsymbol{\vartheta}$ appears in the underlying linear evolution system, the maximiser (4.93) becomes identical to that given in the SPDE in the linear observations case. Specifically, if the prior over $\boldsymbol{\vartheta}$ is given by $\boldsymbol{\vartheta} \sim \mathcal{N}_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}_p, \boldsymbol{\Sigma}_{\boldsymbol{\vartheta},p})$,

$$\tilde{p}(\boldsymbol{\vartheta})^{(i+1)} = \mathcal{N}_{\boldsymbol{\vartheta}}(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}), \tag{4.98}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}} = \left(\sigma_w^{-2}\boldsymbol{\Upsilon} + \boldsymbol{\Sigma}_{\boldsymbol{\vartheta},p}^{-1}\right)^{-1}, \tag{4.99}$$

$$\hat{\boldsymbol{\vartheta}} = \boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}\left(\sigma_w^{-2}\boldsymbol{\upsilon} + \boldsymbol{\Sigma}_{\boldsymbol{\vartheta},p}^{-1}\hat{\boldsymbol{\vartheta}}_p\right), \tag{4.100}$$

where $\boldsymbol{\upsilon}$ and $\boldsymbol{\Upsilon}$ are given in (3.38) and (3.39) respectively. The VBEM algorithm for SPDEs observed through a point process is given in Algorithm 4.1

**Algorithm 4.1** The VBEM algorithm for SPDEs with point process observations

**Input:** Data set $\mathcal{Y}$, parameters $\widetilde{\boldsymbol{Q}}, \{\boldsymbol{V}_i\}_{i=1}^{d}$, $\boldsymbol{\Psi}_{\boldsymbol{x}}, \mu, \bar{\beta}, \sigma_w^{-2}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$ and initial parameter distribution $\tilde{p}(\boldsymbol{\vartheta})^{(0)}$.

```
$i = 0$
do
    Run Algorithm A.3 with $\tilde{p}(\boldsymbol{\vartheta}) = \tilde{p}(\boldsymbol{\vartheta})^{(i)}$        VBE-step
    Evaluate $\tilde{p}(\boldsymbol{\vartheta})^{(i+1)}$ from (4.99) and (4.100)        VBM-step
    $i = i + 1$
until $||\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})^{(i)}}[\boldsymbol{\vartheta}] - \mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta})^{(i-1)}}[\boldsymbol{\vartheta}]|| < \epsilon_1$
```

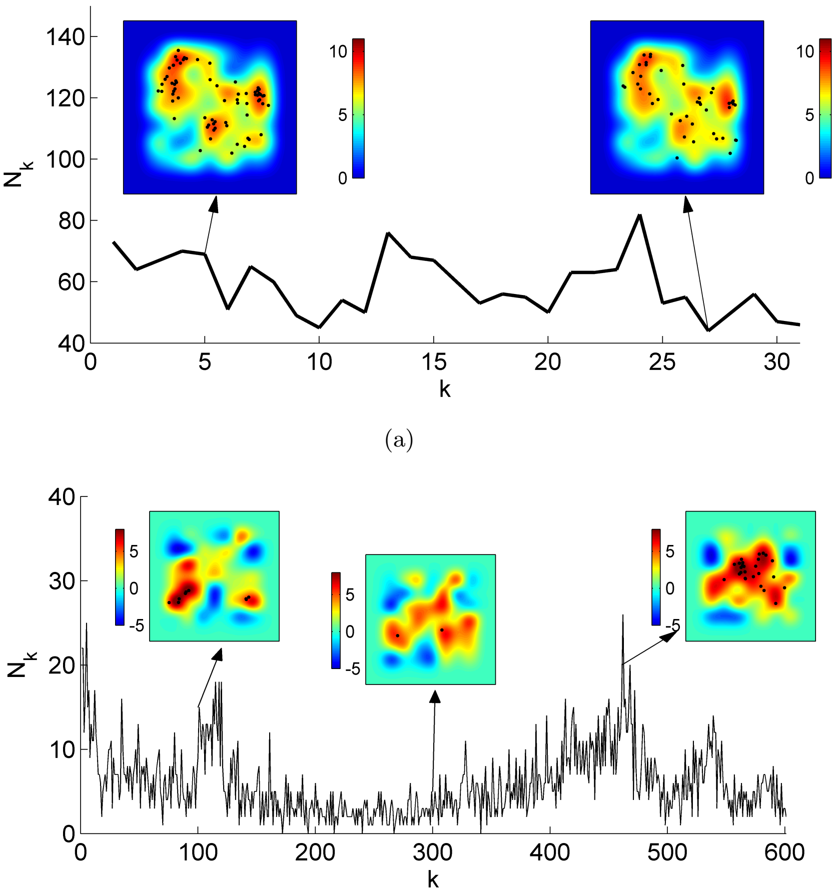
**Output:** $\{\hat{\boldsymbol{x}}_{k|K}, \boldsymbol{\Sigma}_{k|K}\}_{k=0}^{K}, \tilde{p}(\boldsymbol{\vartheta})^{(i)}$.

## 4.5 Case study: the stochastic heat equation

Consider the stochastic heat equation with operator $\mathcal{A}(\cdot) = D\Delta(\cdot), D \in \mathbb{R}^+$ being the thermal diffusivity and $\Delta$ denoting the Laplacian. Assume that a set of events are a realisation of a Poisson process fully defined through its CIF $\lambda(\boldsymbol{s}, t)$ which in turn is conditional on the underlying heat equation through $z(\boldsymbol{s}, t)$. The problem is to find the diffusivity constant $D$ from the discrete events which appear in space and time. Such a case study lends itself to several applications such as i) epidemiology, where typically, instead of the actual *raw* physical locations of reports, a summarised infective population concentration is used in a reaction-diffusion like model [197] and ii) ecology, where the spread of an animal or plant species (again using reaction-diffusion type models) is typically studied by binning the number of reports at discrete location points and carrying out the inference over a relatively coarse grid [27]. In what follows a rigorous method which maintains the precise locations of the individual reports and which adheres to a continuous-space representation is demonstrated to be a suitable tool for the analysis of such systems.

### 4.5.1 Simulation setup

Let $\mathcal{O} = [0, 30] \times [0, 30] \subset \mathbb{R}^2$. A mesh of $n^{sim} = 49$ basis function of the form (3.20) with $\tau^{sim} = 1.05$ were equally spaced on $\mathcal{O}$ satisfying the Dirichlet boundary conditions. In a geographical context, these boundary conditions have a meaningful relevance, such



(a)

(b)

Figure 4.8: Number of events $N_k$ at each time point $k$ for the two case studies under consideration: the stochastic heat equation observed through point processes (a) with a dominating initial condition where the diffusion action is clearly observed and (b) with an initial condition as the process limiting distribution. The insets show the spatial points (black points) and the underlying field $z_k$ (coloured map) at different time instances.

as a coastline over which an epidemic cannot spread, or a country's boundaries beyond which any reported events are irrelevant (see Chapter 6 and the discussion of Illian and Simpson in [20]). Simulation parameters were set as $\Delta_t^{sim} = 0.01$, $\mu = -4$, $\bar{\beta} = 1$, $Qu = \int_{\mathcal{O}} k_Q(\boldsymbol{s} - \boldsymbol{r})u(\boldsymbol{r})\mathrm{d}\boldsymbol{r}$ and with $k_Q(\boldsymbol{s}) = \exp(-\boldsymbol{s}^T\boldsymbol{s}/4)$. The diffusion constant chosen was $D = 10$ so that $\vartheta = 10$. The example in Section 3.1.1 shows how spatially varying parameters can be dealt with in the analysis if desired.

Simulation of the point process was carried out by the method of *thinning* [198]. Essentially this involves generating a homogeneous Poisson process at each time frame with the maximum intensity $\lambda_k^* = \sup \lambda_k(\boldsymbol{s})$ and then randomly removing points in accordance with $\lambda_k(\boldsymbol{s})$. See [32] for an elegant description of the procedure. Two case studies are considered here:

1. A strong initial condition dominates the underlying field so that the process is first-order non-stationary with, however, a substantial amount of events observed bearing considerable information on $\vartheta$ (with $K = 30$, $\sigma_w = 0.5$). The true initial states for this case study were initialised as $\boldsymbol{x}_0 \sim \mathcal{N}_{\boldsymbol{x}_0}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with $\boldsymbol{\mu}_0 \sim \mathcal{U}(0, 7)$, $\mathcal{U}(\cdot, \cdot)$ being the uniform distribution in a given interval, and $\boldsymbol{\Sigma}_0 = 50\boldsymbol{I}$.

2. A stationary spatiotemporal process where the diffusion constant has to be estimated from just an average of about 6 events per frame. More data is needed for reliable estimation in these scenarios (in this case $K = 600$, $\sigma_w = 3$). The initial estimated state for this case study was set to $\boldsymbol{x}_0 \sim \mathcal{N}_{\boldsymbol{x}_0}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with $\boldsymbol{\mu}_0 = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_0 = 50\boldsymbol{I}$.

Figure 4.8 shows two realisations from these two case studies. Inference in the former case is obviously easier from an information perspective, although it will be seen that the algorithm performs well in both cases.

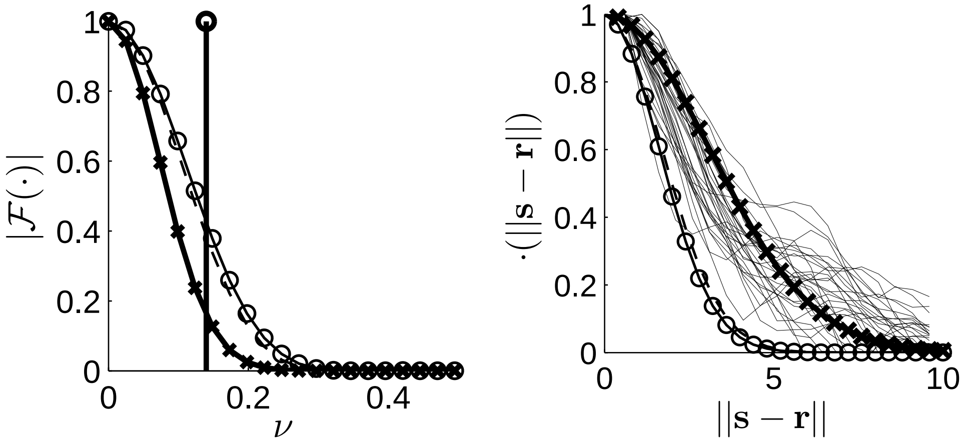
### 4.5.2 Basis function placement

Consider the first data set of Figure 4.8a. The corresponding nonparametric estimators of the PACF (Figure 4.9b) clearly show that the cluster sizes correspond to a field frequency response of approximately $\nu_c = 0.14$ cycles per unit (Figure 4.9a). The methodology in Section 4.4.1 can then easily be used to propose local GRBFs with parameter $\tau = 1.1$. The resulting basis turns out to be quite close to the basis used to generate the data. Moreover note how the selected basis function is naturally found to be narrower than the log auto-correlation function in order to adequately represent cluster hot spots.

In Section 3.3 it was seen that a different form of basis from the simulation always results in some form of bias. Since the proposed basis functions turned out to be so close to those used in the simulation model, and since a discussion of the bias introduced by finite dimensional reduction was already presented in Chapter 3, here the exact simulation representation will be used for modelling (which corresponds to an oversampling parameter of $\alpha_0 = 1.2$). A demonstration of basis function selection from point processes will be employed with the real application data in Chapter 6.

### 4.5.3 Results

A zero-centred Gaussian distribution with variance $\sigma_\vartheta^2 = 1,000$ was used for a non-informative parameter prior. The VB algorithm was terminated with the stopping con-



(a)

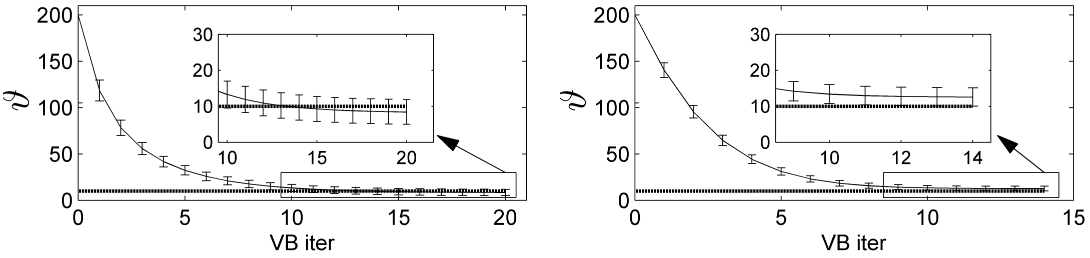
(b)

Figure 4.9: Frequency and spatial analysis of spatiotemporal point process systems using the PACF. (a) Normalised absolute frequency response of the underlying field computed from the log PACF (cross), chosen support of field frequency content (vertical line) and proposed basis function frequency response (circle). (b) Nonparametric estimation of $\ln g_{k,k}$ at all time points (grey solid lines), a GRBF fitted to the mean over all time points (cross) and the proposed basis function (circle). The curves pertaining to the basis functions used to generate the data are also given for comparison (dashed line).

dition $|\mathbb{E}_{\tilde{p}(\vartheta)^{(i)}}[\vartheta] - \mathbb{E}_{\tilde{p}(\vartheta)^{(i-1)}}[\vartheta]| < 0.1$ where $i$ denotes the VB iteration number. SCG was carried out using the MATLAB package NETLAB [199].

Figure 4.10 shows the successful convergence of the VB algorithm for the data sets shown in Figure 4.8a and Figure 4.8b after 20 and 14 iterations respectively. Fifty MC runs were carried out under the same simulation conditions of Figure 4.8a but with $D = 50$. The large parameter uncertainty, as seen in Figure 4.11, comes as no surprise given the small $K$. Interestingly the distribution of the mean is not normal around the true value in this case (Figure 4.11a), and is slightly positively skewed in a way similar to linear systems with high observation noise (see Footnote 9 on pg. 83). In this case the true parameter lay consistently within the three-sigma uncertainty intervals (Figure 4.11b) however, recall from Section 3.4, that this need not be true in general.

The number of iterations required for convergence highly depends on the value of the true parameter $\vartheta$ and a lower diffusion constant tends to speed up convergence (see Figure 4.12). As alluded to in Figure 3.11 in Section 3.4, there are also other factors affecting the number of iterations required for convergence, such as the initial parameter estimates, noise levels and the length of the data set (longer data sets require less iterations for convergence). As a result of the required nonlinear optimisation through



(a)

(b)

Figure 4.10: Estimated parameter posterior mean and three-sigma confidence intervals with increasing VB iteration number for (a) the data set of Figure 4.8a and (b) the data set of Figure 4.8b. The true value to which the means converge to are given by the horizontal dotted lines.

gradient descent, the computational time required by the VB algorithm with point process observations is considerably more than that required with continuous observations. In all cases shown in Figure 4.12, for $n = 49$, inference took on the order of two hours to complete on a standard PC;[10] a similar problem with linear observations would take a few minutes at most.

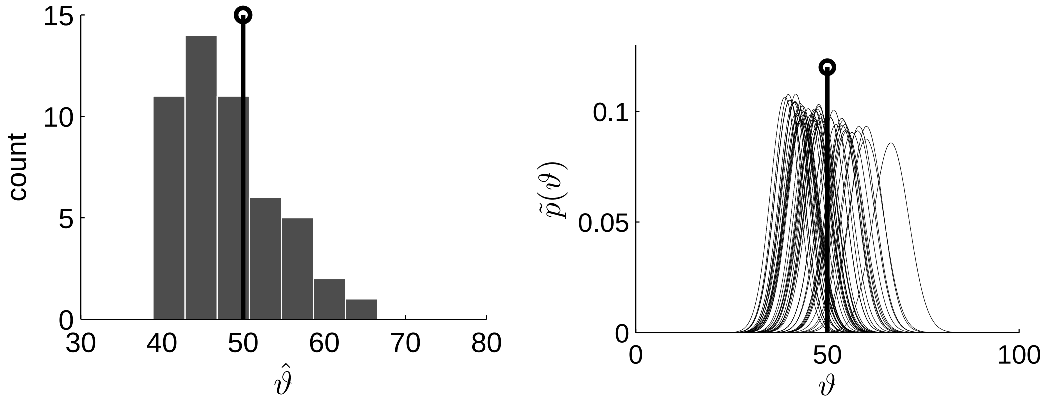
High quality estimation of the underlying spatiotemporal field is shown in Figure 4.13. Note how the smoothing effect of the two-filter smoother has the effect of correctly estimating the field even when only a few events are observed at any point in time. The right panel in each subplot depicts the uncertainty of the field estimate; as a result of the chosen exponential form of $\lambda_k(\mathbf{s})$ uncertainty is highest in the regions of low intensity; an intensity of $e^{-4}$ and $e^{-8}$ correspond equally well to regions of negligible intensity so that the precise value of $z(\boldsymbol{s})$ in low activity regions is largely uncertain. This does not create any difficulty from an engineering perspective where it is likely that regions of high activity (hot spots) are of the most interest.

## 4.6 Conclusion

This chapter has applied, for the first time, a VB-Laplace approximation for the identification of SPDEs from point process observations. This has resulted in several contributions from an algorithms perspective, namely

- the extension of the basic EM algorithm for point process systems [29] to VBEM for single state multi-channel output systems, showing an increase in estimation

[10] Simulations were carried out on an Intel®Core$^{\mathrm{TM}}$2Duo T5500 @ 1.66GHz personal computer with 2GB of RAM.



(a)

(b)

Figure 4.11: (a) Distribution of the mean parameter estimate $\hat{\boldsymbol{\vartheta}}$ over 50 MC runs. (b) Variational posterior over $\boldsymbol{\vartheta}$ at each MC run. The vertical line in each case indicates the value of the true parameter.

accuracy and significant computational savings when compared to MCMC methods such as Gibbs sampling,

- the establishing of a continuous-space output framework which decouples the discretisation needed for numerical integration from the resolution at which points are being observed,
- the joint field-parameter inference scheme for the identification of SPDEs from point process observations and
- the proposal of the use of frequency analysis methods in spatiotemporal systems for modelling via the use of the PACF.

The proposed method for spatiotemporal identification from point processes is relatively straightforward and is only limited by the size of the set of basis functions, which can range into a few hundreds if required. Its applicability to SPDEs is a major step forward from the treatment of spatially decoupled processes such as that considered in [19], which required a highly computationally intensive Metropolis-Hastings algorithm for parameter inference. The deterministic approach is also highly efficient when compared to the popular Metropolis-adjusted Langevin algorithm (MALA) currently extensively employed in this field [35, 185]. Moreover the VB approach reaps the advantages of Bayesian inference when compared to other deterministic approaches. First, it reduces the errors introduced when considering solely ML estimates (as also noted in [185, Section 8]). Second, it obviates the requirement for bootstrap methods for parameter

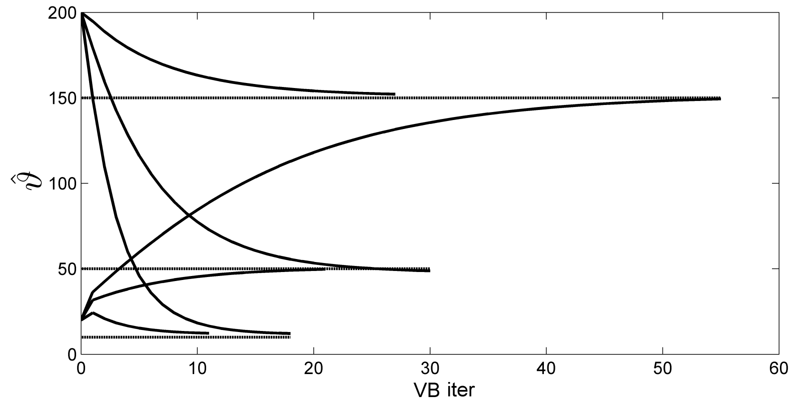


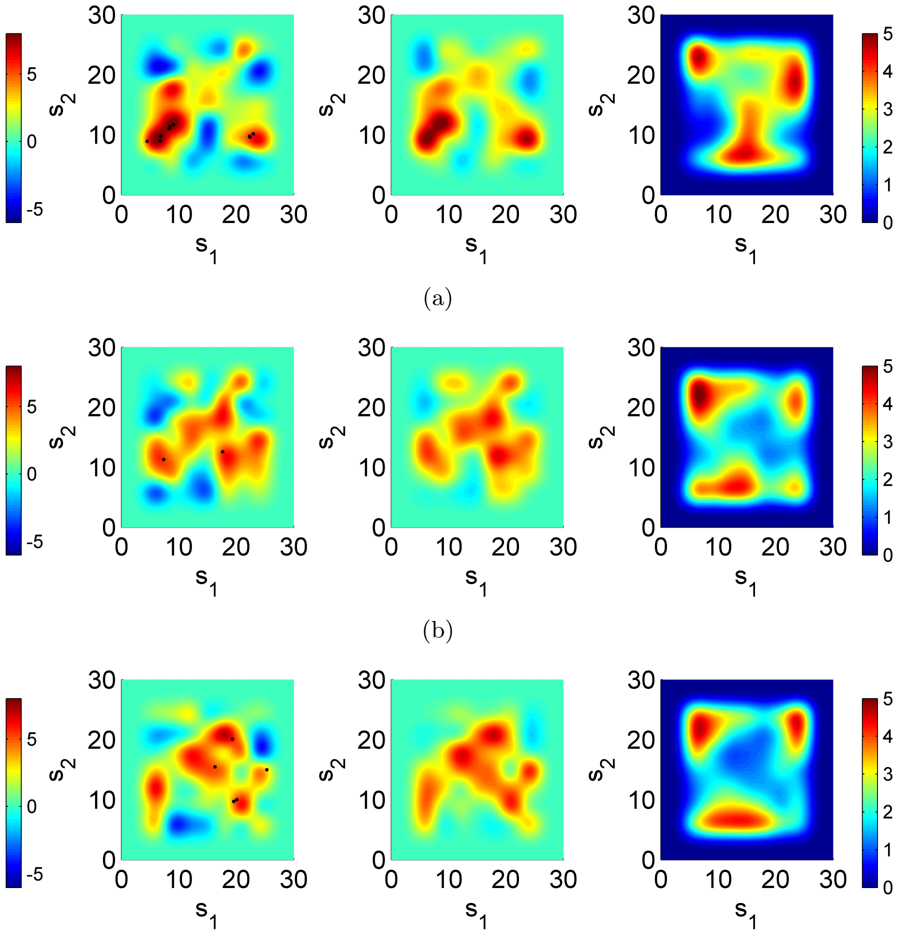
Figure 4.12: Estimated parameter posterior mean $\mathbb{E}_{\tilde{p}(\vartheta)}[\vartheta] = \hat{\vartheta}$ with increasing VB iteration number on data sets of fixed length $K = 31$ with data generated under the simulation conditions used for the data set shown in Figure 4.8a. For clarity confidence intervals are not shown. The true value to which the means converge to are given by the horizontal lines.

uncertainty evaluation as carried out, for instance, in the partial-likelihood approach to point processes [200].

There are several paths for future work, including a more rigorous framework for basis function placement. Throughout the studies it was noted that with spatiotemporal systems with point process observations it is very hard to model any dynamics where there are scarce events (as noted in the case study of Figure 4.8b which required an order of magnitude more in data points for reliable estimation). One method to counter this would be to have a heterogeneous set of local basis functions which have a larger width in less active regions over the whole temporal region, so that the observability of the respective states is guaranteed.

It is noted that the developed state-space Cox process in a spatiotemporal context is also easily adapted to marked/multivariate point processes, where each point has a defining characteristic (such as type of disease incidence in an epidemic or the type of plant which is spreading in ecology [35]). Each mark can have its own state vector and be governed by its own dynamics, and cross-dynamics can be incorporated into the model by the use of an augmented state vector together with an augmented state matrix. The noise statistics also need not be common across the processes through the use of a structured covariance matrix $\widetilde{\boldsymbol{Q}}$.

In light of the objectives set out for in Section 1.3.2, this chapter delivers a practical approach to learn from point process data and provides a discrete-time continuous-space state-space representation of the spatiotemporal system which can readily be used for



(a)

(b)

(c)

Figure 4.13: Field estimation at (a) $k = 100$, (b) $k = 300$ and (c) $k = 500$ depicting the true field $z_k$ overlayed by the observed points (black marks), the estimated field $z_k(\boldsymbol{s}) = \boldsymbol{\phi}^T(\boldsymbol{s})\hat{\boldsymbol{x}}_{k|K}$ (centre panels) and the variance map $\boldsymbol{\phi}^T(\boldsymbol{s})diag(\boldsymbol{\Sigma}_{k|K})$, where $diag(\boldsymbol{\Sigma})$ arranges the diagonal of a matrix $\boldsymbol{\Sigma}$ into a column vector. The colour axis for the left and centre panels corresponds to the colour bar on the left. That of the right panel corresponds to the colour bar on the right.

control purposes. It also adopts fast distributional approximation methods to make the approach amenable to large data sets. Both Chapters 3 and 4 have considered data which is made available a priori to the analyst; the next chapter will extend the work carried out so far to deal with situations where estimation needs to be carried out online.

# Chapter 5

# Online variational inference from spatiotemporal data

Chapters 3 and 4 presented a framework for the offline, or batch, state-parameter estimation of SPDEs from spatiotemporal data available either as continuous readings or as events. The algorithms alternately computed i) the parameters using the smoothed states and ii) the smoothed states using the evaluated parameters. They are optimal in some sense and shown to perform well in a variety of settings. However, whilst being elegant from a statistical viewpoint, as a result of requiring the entire data set for computation, offline methods may be exhausting in terms of memory requirements and computational speed [201]. Moreover, algorithms in this class are less adapted to online scenarios where data is arriving in real-time and where model behaviour is not guaranteed to be time invariant.

Sequential estimators, on the other hand, consider individual measurements as they become available. The information gathered from past data is implicitly stored in the current parameter and state estimates which are updated sequentially with the new data. As a result sequential estimators are seen to require less computational resources [202] than batch estimators. Moreover, while also applicable to offline analysis, they are ideal for online scenarios where data is arriving in real-time and may be adapted to situations where the field dynamics are slowly changing [203]. The conceptual distinction between the two classes of estimators is given in Figure 5.1.

Combined state-parameter sequential estimation in spatiotemporal systems is useful for both types of observation processes considered so far. For instance:

- Epilepsy monitoring [74]: Certain parameter alterations detected in a patient-specific neural field model may be used to anticipate imminent seizures. In this case the observations are electrophysiological data (continuous readings).

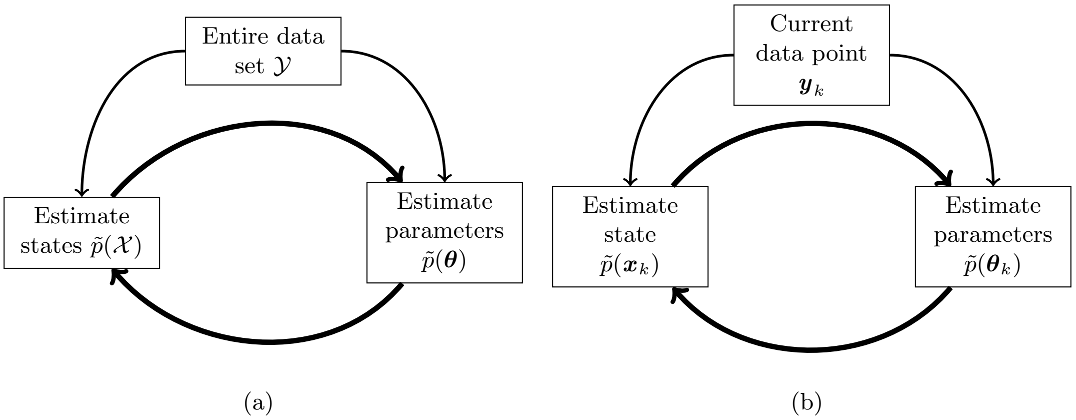

Entire data set $\mathcal{Y}$
Estimate states $\tilde{p}(\mathcal{X})$
Estimate parameters $\tilde{p}(\boldsymbol{\theta})$
(a)
Current data point $\boldsymbol{y}_k$
Estimate state $\tilde{p}(\boldsymbol{x}_k)$
Estimate parameters $\tilde{p}(\boldsymbol{\theta}_k)$
(b)

Figure 5.1: Combined state-parameter estimation. (a) *Batch approach:* The entire data set is used and state estimation over the entire time horizon is alternated with parameter estimation. (b) *Sequential approach:* The current data point $\boldsymbol{y}_k$ is used to give the current state estimate $\boldsymbol{x}_k$ and the current parameter estimate $\boldsymbol{\theta}_k$. Iterations for state and parameter estimation at the current time point $k$ may still be required.

- Disease surveillance [43]: By studying the progression of the logged disease cases, outbreak detection may be anticipated and the ability to counteract the disease spread increased. The outbreak may, for instance, correspond to a shift from stable dynamics to unstable ones. Here the observations are logged disease cases (event-based data).

Other application areas where online estimation might prove profitable in a spatiotemporal context is structural health monitoring [204] and urban development [19].

The main aim of this chapter is to extend the framework considered so far by providing sequential combined estimation of spatiotemporal systems with the use of both sensor data and event-based data in a variational framework. Towards such an aim a brief review of established joint/dual filtering methods together with some preliminaries for this work is given in Section 5.1. A variational dual filter is then derived in Section 5.2. As the variational filter is a key novelty in this chapter its potential is first evaluated on simple temporal systems and then on spatiotemporal systems with continuous observations in Section 5.3. The same approach is then taken for point process observations in Sections 5.4 and 5.5. The chapter concludes in Section 5.6.

## 5.1 Preliminaries

### 5.1.1 Introduction to dual filtering

Combined sequential state-parameter estimation algorithms appear in two flavours, *joint estimation* and *dual estimation*. Both methods require the postulation of an artificial dynamic equation for parameter progression (to be discussed later). In joint estimation, the state vector is augmented with the unknown parameters to form a $n+d$-dimensional state vector which is then updated using the Kalman filter and its nonlinear extensions [205]. However the augmentation does not always yield tractable models and, in addition, has been reported to generate instability [206] particularly in nonlinear dynamic models [207].

Instead of employing a single augmented state vector, a dual filter attempts to estimate the states and parameters sequentially by alternating between two filters running in parallel. The most popular dual filtering technique is the dual Kalman filtering (DKF) technique which considers two standard [208], extended [119], ensemble [207] or unscented [205, 120] Kalman filters running alternately. For nonlinear systems the unscented and ensemble filters appear to be the most promising [209] as they accurately estimate the mean and variance of the state and parameter posterior distributions at each time point. However, the DKF methods only employ a point estimate of the state for computation of the parameter distribution, and likewise, employ only a point estimate of the parameter for computation of the state distribution. The DKF algorithms effectively ignore the computed variances, giving rise to the question as to whether more accurate distributional approximations may be obtained by sequential propagation of higher order moments.

At the other end of the spectrum sequential Monte Carlo (SMC) filters [210, 211], unlike DKF methods, are able to use all the parameter and state moments in the alternating updates. They estimate the full posterior distribution over both the parameters and the states and provide reliable estimates. However, while these methods can often accurately approximate the posterior distribution, their computational cost may easily prove prohibitive in the study of spatiotemporal systems which usually involve a high combined state-space and parameter-space dimensionality [212].

More recently, in what may be seen as an attempt to find a compromise between DKF and SMC methods, online EM filters [201, 213, 214] were developed. Unlike DKF algorithms, these take into account distributions over the states computed in an (online) E-step for the evaluation of a ML point estimate of the parameters computed in the following (online) M-step. The ML estimate is then used for estimation of the states at the next step. The online EM method, to the best of the author's knowledge, has not yet been benchmarked against other methods.

Recent years have seen an emerging interest in VB filters as a further compromise between parameter point estimate filters and online SMC filters [143, 215, 216]. A natural choice when employing VB is to assume temporal independence of the filtered random variables [216] so that, for instance,

$$\begin{aligned} p(\mathcal{X}) &= p(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) \\ &\approx \tilde{p}(\mathcal{X}) \\ &= \tilde{p}(\boldsymbol{x}_0)\tilde{p}(\boldsymbol{x}_1)\ldots\tilde{p}(\boldsymbol{x}_K). \end{aligned} \tag{5.1}$$

This is obviously a strong assumption in dynamic models and particularly in diffusion systems where states are highly correlated in time. This assumption also leads to anomalies such as data/time invariance of covariance matrices [14, Section 7.4].

VB in conjunction with particle filtering has been explored for use with state-space models [143]. The requirement of MC methods for estimation is however still a detriment in high dimensional spatiotemporal settings. The first 'pure' dual VB filter (DVBF) for state-space models appeared in [217] where the states and observation noise precisions were estimated online. The presented filter is nonetheless very tailored for the problem and, in addition, implicitly assumes that parameter estimates at time point $k-1$ are unaffected by the new data point $\boldsymbol{y}_k$ which, as shown in Section 5.2, is not necessarily the case.

The derivation of a novel DVBF is discussed in the next section. It is rigorous, placed in a general setting and elegant in the context of the framework constructed so far which has put a lot of emphasis on variational analysis. The generality contributes to it being easily applied to a variety of systems, including spatiotemporal systems observed using both sensors and events. The DVBF also serves to fill in the gap in the literature of dual filtering, which is currently dominated by Kalman variants and SMC methods. The DVBF may yet prove to be the best compromise between parameter point estimate filters and SMC filters. Since unlike other deterministic approximation methods the DVBF also provides distributions over the unknown parameters, it may be particularly useful in decision making or in adaptive control scenarios.

### 5.1.2 Restricted variational Bayes

In its elementary form, VB is not a suitable tool for sequential estimation since, as explained at the end of this section, all distributions estimated at previous time instants need to be re-estimated under the new, updated distributions given the new data points. To remedy this problem, this section exploits a new class of VB techniques, coined *restricted VB (RVB)* in [218]. For ease of exposition the following theorem is given for two partitions of the set of unknowns, $\mathcal{X}$ and $\boldsymbol{\theta}$.

**Theorem 5.1 (Restricted variational Bayes [14, Section 3.4])** *Let $p(\mathcal{X}, \boldsymbol{\theta}|\mathcal{Y})$ be the true posterior, $\bar{p}(\boldsymbol{\theta})$ a fixed (restricted) distribution over $\boldsymbol{\theta}$ and $\tilde{p}(\mathcal{X}, \boldsymbol{\theta})$ the approximation to the true posterior decomposed in the usual way*

$$\tilde{p}(\mathcal{X}, \boldsymbol{\theta}) \approx \tilde{p}(\mathcal{X})\bar{p}(\boldsymbol{\theta}). \tag{5.2}$$

*Then the maximum of a lower bound of the log evidence is reached for*

$$\tilde{p}(\mathcal{X}) \propto \exp\Big(\mathbb{E}_{\bar{p}(\boldsymbol{\theta})}[\ln p(\mathcal{X}, \boldsymbol{\theta}, \mathcal{Y})]\Big). \tag{5.3}$$

*Proof.* Consider the log-evidence

$$\begin{aligned}\ln p(\mathcal{Y}) &= \ln p(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta}) - \ln p(\mathcal{X}, \boldsymbol{\theta} \mid \mathcal{Y}) \\ &= \mathbb{E}_{\tilde{p}(\mathcal{X})\bar{p}(\boldsymbol{\theta})}\left[\ln \frac{p(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta})}{\tilde{p}(\mathcal{X})\bar{p}(\boldsymbol{\theta})}\right] + \mathbb{E}_{\tilde{p}(\mathcal{X})\bar{p}(\boldsymbol{\theta})}\left[\ln \frac{\tilde{p}(\mathcal{X})\bar{p}(\boldsymbol{\theta})}{p(\mathcal{X}, \boldsymbol{\theta} \mid \mathcal{Y})}\right] \\ &= \mathcal{L}(\tilde{p}(\mathcal{X}), \bar{p}(\boldsymbol{\theta})) + KL(\tilde{p}(\mathcal{X})\bar{p}(\boldsymbol{\theta})||p(\mathcal{X}, \boldsymbol{\theta}|\mathcal{Y})). \end{aligned} \tag{5.4}$$

As in conventional VB it follows that $\ln p(\mathcal{Y}) \geq \mathcal{L}(\tilde{p}(\mathcal{X}), \bar{p}(\boldsymbol{\theta}))$ so that $\mathcal{L}(\cdot, \cdot)$ is a lower bound. The lower bound may be maximised using functional derivatives as in Theorem 2.3, however using solely $\tilde{p}(\mathcal{X})$ since $\bar{p}(\boldsymbol{\theta})$ is restricted, to give (5.3). ■

VB is a general function approximation method where, for instance, the optimal variational posteriors over individual variables are found such that their product approximates the true posterior in an optimal fashion. RVB is similar, except that one or more of the posterior distributions are restricted, assumed to be known (or approximately known) a priori. An advantage is that iterations involving these restricted variables are omitted, resulting in computational savings.[1] Since RVB does not allow modification of $\bar{p}(\boldsymbol{\theta})$ it is sub-optimal, becoming optimal in the variational sense only if $\bar{p}(\boldsymbol{\theta}) = \tilde{p}(\boldsymbol{\theta})$.

**Remark 5.1** *RVB should not be confused with functionally constrained VB where the distribution over the parameters $\boldsymbol{\theta}$ is allowed to change but is of fixed functional form. An example of functionally constrained VB is the VB-Laplace (Section 4.1.2) or the EM algorithm, where the parameter distribution is constrained to be a Gaussian or a delta Dirac respectively.*

In dual filtering it is common to introduce heuristic dynamics for static parameters to aid in the inference procedure. As a result there is a shift from the treatment of $\boldsymbol{\theta}$ to that of $\boldsymbol{\Theta}_k = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$. Denote the $i^{th}$ variational parameter distribution estimated

[1] For the bi-partitioning in Theorem 5.1 no iterations would be required.

at the $i^{th}$ time instant as $\tilde{p}(\boldsymbol{\theta}_i|\mathcal{Y}_i)$, where $\mathcal{Y}_i = \boldsymbol{y}_{1:i}$, and that estimated at the $k^{th}$ time instant as $\tilde{p}(\boldsymbol{\theta}_i|\mathcal{Y}_k)$. Since $\tilde{p}(\boldsymbol{\theta}_i|\mathcal{Y}_k), i < k$ is in general not equal to $\tilde{p}(\boldsymbol{\theta}_i|\mathcal{Y}_i)$, with the use of standard estimation sequential estimation (without backtracking), re-estimating the parameters at the previous time points is not possible. Hence RVB is an integral part of the DVBF derived in the next section; by additionally constraining the full (temporal) variational parameter posterior, one can obtain a recursive algorithm in which only the last two variational posteriors of the states and parameters, respectively, are coupled. It is also seen that in state-space models, constraining is only required on the parameters; the temporal trajectory of the states remains unchanged.

## 5.2 The dual VB filter

### 5.2.1 Factorisation of the joint posterior distribution

Consider the following general nonlinear, non-Gaussian dynamic state-space model,

$$\boldsymbol{x}_k = F(\boldsymbol{\theta}_k, \boldsymbol{x}_{k-1}, \boldsymbol{w}_k), \tag{5.5}$$

$$\boldsymbol{y}_k = G(\boldsymbol{\theta}_k, \boldsymbol{x}_k, \boldsymbol{v}_k), \tag{5.6}$$

where $\boldsymbol{x}_k \in \mathbb{R}^n$ and $\boldsymbol{y}_k \in \mathbb{R}^m$ are the state and observation vectors at $k$ respectively, with $n, m \in \mathbb{Z}^+$. $F(\cdot)$ and $G(\cdot)$ are smooth functions and $\boldsymbol{w}_k \in \mathbb{R}^n$ and $\boldsymbol{v}_k \in \mathbb{R}^m$ are noise terms distributed as $\boldsymbol{w}_k \sim \mathcal{N}_{\boldsymbol{w}_k}(\mathbf{0}, \boldsymbol{\Sigma}_w)$ and $\boldsymbol{v}_k \sim \mathcal{N}_{\boldsymbol{v}_k}(\mathbf{0}, \boldsymbol{\Sigma}_v)$, where $\boldsymbol{\Sigma}_w \in \mathbb{R}^{n\times n}$ and $\boldsymbol{\Sigma}_v \in \mathbb{R}^{m\times m}$. The unknown parameter vector $\boldsymbol{\theta}_k \in \mathbb{R}^d, d \in \mathbb{Z}^+$ is assumed to evolve as

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \boldsymbol{e}_{k-1}, \tag{5.7}$$

where $\boldsymbol{e}_{k-1} \in \mathbb{R}^d$ is additive white Gaussian noise with time-varying covariance $\boldsymbol{\Sigma}^e_{k-1} \in \mathbb{R}^{d\times d}$.

Let $\mathcal{X}_k, \mathcal{Y}_k$ be the set of states and observed data set up to time $k$ respectively i.e. $\mathcal{X}_k = \{\boldsymbol{x}_i\}_{i=0}^k = \boldsymbol{x}_{0:k}$ and $\mathcal{Y}_k = \{\boldsymbol{y}_i\}_{i=1}^k = \boldsymbol{y}_{1:k}$. Similarly, let $\boldsymbol{\Theta}_k = \{\boldsymbol{\theta}_i\}_{i=1}^k$. For this problem, at time $k$ the true posterior distribution is given by $p(\mathcal{X}_k, \boldsymbol{\Theta}_k|\mathcal{Y}_k)$. It is desired, as a result of the Markovian dynamics of $\boldsymbol{x}_k$, to find a suitable approximate factorised approximation which preserves the conditional dependency between the states. Preservation of the temporal dependencies of the parameters in time is deemed less important as the dynamics introduced are not physically representative. Hence, a suitable approximation is given as

$$\tilde{p}(\mathcal{X}_k, \mathbf{\Theta}_k) \approx \tilde{p}(\mathcal{X}_k) \prod_{j=1}^{k} \tilde{p}(\boldsymbol{\theta}_j) = \tilde{p}(\mathcal{X}_k)\tilde{p}(\mathbf{\Theta}_k), \tag{5.8}$$

where $\tilde{p}(\mathbf{\Theta}_k)$ is the product of the variational posterior distributions estimated using the data up to the current time instant $k$.

Assume further that $\boldsymbol{\theta}_k = \{\theta_k^1, \theta_k^2, \ldots, \theta_k^l\}$, which constitutes a set of conditionally independent parameters under the variational approximation. Then, expressions for the variational state posterior distribution $\tilde{p}(\mathcal{X}_k)$ and each variational parameter posterior distribution at the final time instant $\tilde{p}(\theta_k^i)$ which maximise the lower bound are given by (Theorem 2.3)

$$\tilde{p}(\mathcal{X}_k) \propto \exp(\mathbb{E}_{\tilde{p}(\mathbf{\Theta}_k)}[\ln p(\mathcal{X}_k, \mathcal{Y}_k, \mathbf{\Theta}_k)]), \tag{5.9}$$

$$\tilde{p}(\theta_k^i) \propto \exp(\mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\mathbf{\Theta}_k^{/\theta_k^i})}[\ln p(\mathcal{X}_k, \mathcal{Y}_k, \mathbf{\Theta}_k)]), \qquad i = 1 \ldots l. \tag{5.10}$$

Since $\tilde{p}(\mathcal{X}_k)$ needs to be re-estimated under the updated sequence of parameter distributions $\{\tilde{p}(\theta_j^i)\}_{i,j=1}^{l,k}$, without any further restrictions (as a result of the required backtracking for parameter estimation) it is very difficult to find a recursive solution to (5.9) and (5.10) for each new data. RVB is therefore now employed to further restrict variational posteriors so that they are conditional only up to the time instant in which they were estimated. Rendering explicit the subset of the data with which the variational posterior was computed, $\tilde{p}(\mathbf{\Theta}_k)$ is now given as

$$\begin{aligned}\tilde{p}(\mathbf{\Theta}_k) &= \tilde{p}(\boldsymbol{\theta}_k|\mathcal{Y}_k) \prod_{j=1}^{k-1} \bar{p}(\boldsymbol{\theta}_j|\mathcal{Y}_j) \\ &= \tilde{p}(\boldsymbol{\theta}_k|\mathcal{Y}_k)\bar{p}(\mathbf{\Theta}_{k-1}).\end{aligned} \tag{5.11}$$

At each time step, the distributions $\tilde{p}(\mathcal{X}_k)$ and $\tilde{p}(\boldsymbol{\theta}_k) = \tilde{p}(\theta_k^1)\tilde{p}(\theta_k^2)\ldots\tilde{p}(\theta_k^l)$ are the usual variational posteriors, whilst

$$\bar{p}(\mathbf{\Theta}_{k-1}) = \prod_{j=1}^{k-1}\prod_{i=1}^{l} \bar{p}(\theta_j^i). \tag{5.12}$$

are the restricted variational posteriors.

### 5.2.2 Sequential estimation

The key insight of the DVBF is that, through the use of RVB, an appropriate recursive estimation of the approximate full joint posterior is made possible. This is shown in the following theorem.

**Theorem 5.2** *For the state-space equations (5.5) and (5.6), given the factorisation (5.8), the restriction (5.12) and the maximisers (5.9) and (5.10), the recursive updates for the state distribution $\tilde{p}(\mathcal{X}_k)$ and each $\tilde{p}(\theta^i_k), i = 1 \ldots l$, are given by*

$$\tilde{p}(\boldsymbol{x}_k) \propto \int \tilde{p}(\boldsymbol{x}_{k-1}) \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\ln p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{\theta}_k) p(\boldsymbol{y}_k|\boldsymbol{x}_k, \boldsymbol{\theta}_k)])\, d\boldsymbol{x}_{k-1}, \tag{5.13}$$

$$\begin{aligned}\tilde{p}(\theta^i_k) \propto \exp(\mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\boldsymbol{\theta}^{/i}_k)}&[\ln p(\boldsymbol{y}_k|\boldsymbol{x}_k, \boldsymbol{\theta}_k) p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{\theta}_k)]) \\ &\times \ \exp(\mathbb{E}_{\bar{p}(\theta^i_{k-1})}[\ln p(\theta^i_k|\theta^i_{k-1})]), \quad i = 1 \ldots l.\end{aligned} \tag{5.14}$$

*Proof.* Consider the variational approximation of the state marginal given by

$$\begin{aligned}\tilde{p}(\boldsymbol{x}_k) &\propto \int \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta}_k)}[\ln p(\mathcal{X}_k, \boldsymbol{\Theta}_k, \mathcal{Y}_k)])\mathrm{d}\mathcal{X}_{k-1} \\ &= \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\ln p(\boldsymbol{y}_k|\boldsymbol{x}_k, \boldsymbol{\theta}_k)]) \int \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta}_k)}[\ln\{p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{\theta}_k) \\ &\qquad \times p(\mathcal{X}_{k-1}, \boldsymbol{\Theta}_k, \mathcal{Y}_{k-1})\}])\mathrm{d}\mathcal{X}_{k-1}.\end{aligned} \tag{5.15}$$

The second quantity in the integrand can also be expanded, and by treating the conditional parameter distributions as constants relative to the distribution of interest, it can be shown that

$$\begin{aligned}\tilde{p}(\boldsymbol{x}_k) \propto &\exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\ln p(\boldsymbol{y}_k|\boldsymbol{x}_k, \boldsymbol{\theta}_k)]) \int \Bigg( \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\ln p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{\theta}_k)]) \\ &\times \Bigg[ \exp(\mathbb{E}_{\bar{p}(\boldsymbol{\theta}_{k-1})}[\ln p(\boldsymbol{y}_{k-1}|\boldsymbol{x}_{k-1}, \boldsymbol{\theta}_{k-1})]) \int \exp(\mathbb{E}_{\bar{p}(\boldsymbol{\theta}_{k-1})}[\ln p(\boldsymbol{x}_{k-1}|\boldsymbol{x}_{k-2}, \boldsymbol{\theta}_{k-1})]) \\ &\times \exp(\mathbb{E}_{\bar{p}(\boldsymbol{\Theta}_{k-1})}[\ln p(\mathcal{X}_{k-2}, \boldsymbol{\Theta}_{k-1}, \mathcal{Y}_{k-2})])\mathrm{d}\mathcal{X}_{k-2} \Bigg] \Bigg) \mathrm{d}\boldsymbol{x}_{k-1}.\end{aligned} \tag{5.16}$$

Since by RVB the approximate parameter posteriors have been restricted to be conditional on the data up to the instant in which they were estimated, the distributions of the parameters do not need to be recomputed using the latest data which is available.

In particular for any function $\psi(\cdot)$

$$\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta}_k)}[\psi(\boldsymbol{\theta}_{k-1})] = \mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_{k-1})}[\psi(\boldsymbol{\theta}_{k-1})], \tag{5.17}$$

which was computed at the previous time step. Hence, in comparison to (5.15), it is clear that the terms in the square brackets of (5.16) constitute the exact variational posterior of the state at the previous time instant to give (5.13.) Equation (5.14) follows by application of the chain rule on (5.10), where the joint $p(\mathcal{X}_{k-1}, \boldsymbol{\Theta}_k^{/\theta_k^i}, \mathcal{Y}_{k-1})$ is constant relative to the distribution of interest. ■

The above does not constitute an online algorithm in the strictest sense since (5.13) and (5.14) are evidently coupled, and, as in the offline case, some form of iterations between the solutions is required for convergence. However, iterations are required only between the posteriors at the last time instant, making the algorithm fast and efficient, and in practice few iterations often suffice.

## 5.3 Dual filtering from continuous observations

In the author's work [56] it is shown that with continuous observations the DVBF performs similarly to Rao-Blackwellised particle filters (PF) [121], while exhibiting considerable computational savings. The scope of this section is to first extend the analysis and show that the DVBF is more accurate, in some sense, than the DKF and then apply the DVBF to estimate *online* a spatiotemporally varying parameter in a system governed by a SPDE.

### 5.3.1 Comparison with DKF methods

Consider the usual linear state-space model

$$\boldsymbol{x}_k = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x}_{k-1} + \boldsymbol{w}_k, \tag{5.18}$$

$$\boldsymbol{y}_k = \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{v}_k, \tag{5.19}$$

where the matrices

$$\boldsymbol{A} = \begin{bmatrix} 0.1 & 0.5 \\ 0.5 & 0.1 \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} 10 & 5 \\ 1 & 10 \end{bmatrix}, \tag{5.20}$$

and the disturbances $\boldsymbol{\Sigma}_v = \sigma_v^2\boldsymbol{I}$, $\boldsymbol{\Sigma}_w = 0.1\boldsymbol{I}$. The transition matrix $\boldsymbol{A} = [a_{i,j}]_{i,j=1}^2$ is assumed to be fully parameterised by the parameter vector $\boldsymbol{\theta}$ with $\theta_1 = a_{1,1}, \theta_2 = a_{2,1}, \theta_3 = a_{1,2}$ and $\theta_4 = a_{2,2}$. The aim in dual filtering here is to find recursive estimates of $\boldsymbol{x}_k$ and $\boldsymbol{\theta}_k$ at each new data point $\boldsymbol{y}_k$.

**DKF implementation:** For estimation purposes in the DKF the parameter vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3, \theta_4]^T$ is assumed to follow the dynamic model

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \boldsymbol{e}_{k-1}, \tag{5.21}$$

where $\boldsymbol{e}_{k-1} \sim \mathcal{N}_{\boldsymbol{e}_{k-1}}(\boldsymbol{0}, (\lambda^{-1} - 1)\boldsymbol{\Sigma}_{k-1|k-1}^{\boldsymbol{\theta}})$, $\lambda$ is a user-defined forgetting factor (to be set later) and $\boldsymbol{\Sigma}_{k-1|k-1}^{\boldsymbol{\theta}}$ is the posterior variance of the parameters at $k-1$ [119]. The forgetting factor $\lambda \in (0, 1]$ serves to widen the variance of the predictive distribution of $\boldsymbol{\theta}_k$ with respect to the posterior variance of $\boldsymbol{\theta}_{k-1}$, thus aiding convergence and allowing for parameter tracking.[2] In turn the state transition equation may be re-written as

$$\boldsymbol{x}_k = \boldsymbol{A}(\boldsymbol{\theta}_k)\boldsymbol{x}_{k-1} + \boldsymbol{w}_k, \tag{5.22}$$

where $\boldsymbol{x}_k$ and $\boldsymbol{\theta}_k$ need to be found at the arrival of every new data point.

For this simple problem, the optimal DKF method consists of two standard Kalman filters running in parallel,[3] where the dynamic equations for the state and parameter filter are given respectively by

$$\left.\begin{aligned} \boldsymbol{x}_k &= \boldsymbol{A}(\hat{\boldsymbol{\theta}}_{k|k-1})\boldsymbol{x}_{k-1} + \boldsymbol{w}_k \\ \boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{v}_k \end{aligned}\right\} \begin{array}{c}\text{State-space}\\ \text{model,}\end{array} \qquad \left.\begin{aligned} \boldsymbol{\theta}_k &= \boldsymbol{\theta}_{k-1} + \boldsymbol{e}_{k-1} \\ \hat{\boldsymbol{x}}_{k|k} &= \boldsymbol{C}_k^{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{k-1|k-1})\boldsymbol{\theta}_k + \boldsymbol{w}_k \end{aligned}\right\} \begin{array}{c}\text{Parameter-}\\ \text{space model.}\end{array}$$

The parameters are observed through the state estimates in [119]

$$\boldsymbol{C}_k^{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_{k-1|k-1}) = \begin{bmatrix} \hat{x}_{1,k-1|k-1} & 0 & \hat{x}_{2,k-1|k-1} & 0 \\ 0 & \hat{x}_{1,k-1|k-1} & 0 & \hat{x}_{2,k-1|k-1} \end{bmatrix}. \tag{5.23}$$

The DKF first performs a Kalman filtering step to find an optimal estimate of $\boldsymbol{x}_k$, $\hat{\boldsymbol{x}}_{k|k}$, using a predictor value of $\boldsymbol{\theta}_k$ which by (5.21) is equal to the mean parameter value at $k-1$, $\hat{\boldsymbol{\theta}}_{k-1|k-1}$. The parameter correction is subsequently found using the posterior state estimates in the parameter observation equation. Complete implementation details may be found in [119].

[2] In fact $p(\boldsymbol{\theta}_k|\mathcal{Y}_{k-1}) = \mathcal{N}_{\boldsymbol{\theta}_k}(\boldsymbol{0}, \lambda^{-1}\boldsymbol{\Sigma}_{k-1|k-1}^{\boldsymbol{\theta}})$ [119].

[3] Note that this does not imply that the DKF gives the optimal solution to the dual estimation problem.

**DVBF implementation:** Since in (5.14) the parameter dynamics appear directly (instead of the predictive distribution), $\boldsymbol{\theta}$ is assumed to follow the same model (5.21) but with $\boldsymbol{e}_{k-1} \sim \mathcal{N}_{e_{k-1}}(\mathbf{0}, \lambda^{-1}\boldsymbol{\Sigma}^{\boldsymbol{\theta}}_{k-1|k-1})$. This approach ensures that the learning rates of the DKF and the DVBF are the same. Although this representation of the parameter dynamics will also be used throughout this chapter, this need not be the case in general.

Applying Theorem 5.2 to this problem yields the state distribution

$$\tilde{p}(\boldsymbol{x}_k) \propto \mathcal{N}_{\boldsymbol{x}_k}(\hat{\boldsymbol{x}}_{k|k}, \boldsymbol{\Sigma}_{k|k}), \tag{5.24}$$

where

$$\boldsymbol{\Sigma}^{-1}_{k|k} = \boldsymbol{\Sigma}^{-1}_w + \boldsymbol{C}^T\boldsymbol{\Sigma}^{-1}_v\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\boldsymbol{A}_k]\boldsymbol{\Sigma}^*_{k-1}\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\boldsymbol{A}^T_k]\boldsymbol{\Sigma}^{-1}_w, \tag{5.25}$$

$$\hat{\boldsymbol{x}}_{k|k} = \boldsymbol{\Sigma}^{-1}_{k|k}[\boldsymbol{C}^T\boldsymbol{\Sigma}^{-1}_v\boldsymbol{y}_k + \boldsymbol{\Sigma}^{-1}_w\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\boldsymbol{A}_k]\boldsymbol{\Sigma}^*_{k-1}\boldsymbol{\Sigma}^{-1}_{k-1|k-1}\hat{\boldsymbol{x}}_{k-1|k-1}], \tag{5.26}$$

and where

$$\boldsymbol{\Sigma}^*_{k-1} = (\boldsymbol{\Sigma}^{-1}_{k-1|k-1} + \mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\boldsymbol{A}^T_k\boldsymbol{\Sigma}^{-1}_w\boldsymbol{A}_k])^{-1}. \tag{5.27}$$

Note that these are identical to the equations obtained in the forward pass of the offline VB algorithm (see Algorithm A.2), but with expectations taken with respect to $\tilde{p}(\boldsymbol{\theta}_k)$ instead of $\tilde{p}(\boldsymbol{\theta})$. This will always be the case with the DVBF. The state update requires the evaluation of the quantity $\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\boldsymbol{A}^T_k\boldsymbol{\Sigma}^{-1}_w\boldsymbol{A}_k]$. This is quite involved for the parameterisation under consideration and technical details are given in Appendix D.1.

The corresponding parameter update is given by

$$\tilde{p}(\boldsymbol{\theta}_k) = \mathcal{N}_{\boldsymbol{\theta}_k}(\hat{\boldsymbol{\theta}}_{k|k}, \boldsymbol{\Sigma}^{\boldsymbol{\theta}}_{k|k}), \tag{5.28}$$

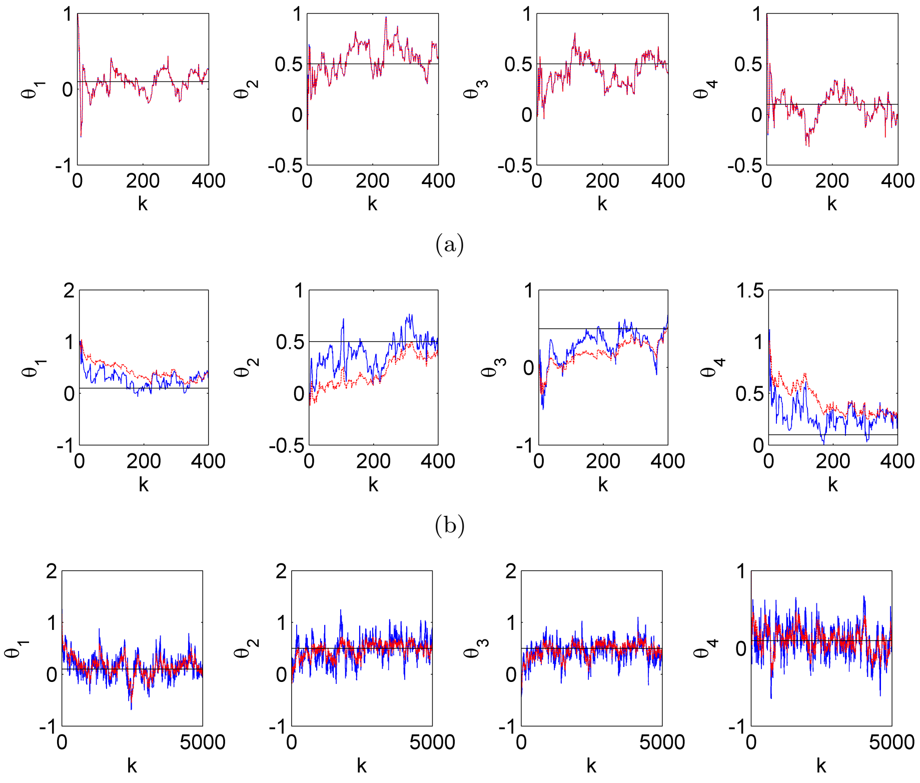
where

$$\boldsymbol{\Sigma}^{\boldsymbol{\theta}^{-1}}_{k|k} = (\boldsymbol{\Lambda}_{k-1} \otimes \boldsymbol{\Sigma}^{-1}_w) + \lambda\boldsymbol{\Sigma}^{\boldsymbol{\theta}^{-1}}_{k-1|k-1}, \tag{5.29}$$

$$\hat{\boldsymbol{\theta}}_{k|k} = \boldsymbol{\Sigma}^{\boldsymbol{\theta}}_k[\lambda\boldsymbol{\Sigma}^{\boldsymbol{\theta}^{-1}}_{k-1|k-1}\hat{\boldsymbol{\theta}}_{k-1|k-1} + (\boldsymbol{\Gamma}^T_k \otimes \boldsymbol{\Sigma}^{-1}_w)vec(\boldsymbol{I})], \tag{5.30}$$

where $\otimes$ is the Kronecker product, $vec(\cdot)$ is the vectorisation operator and $\boldsymbol{I}$ is the identity matrix. Recall that in (5.29), $\boldsymbol{\Lambda}_{k-1} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{\Sigma}^{-1}_w \in \mathbb{R}^{n\times n}$ so that $\boldsymbol{\Sigma}^{\boldsymbol{\theta}^{-1}}_{k|k} \in \mathbb{R}^{n^2\times n^2}$.

The derivation for (5.24)-(5.27) follows on the same lines as that for offline VBEM [176] and is therefore omitted. The derivation for (5.28)-(5.30) is non-standard and given in Appendix D.2. The quantities $\boldsymbol{\Lambda}_{k-1}$ and $\boldsymbol{\Gamma}_k$ are the same as those in (3.42) applied



(a)

(b)

(c)

Figure 5.2: Dual state-parameter sequential estimation of the system (5.18), (5.19) using the DKF method (blue) and the DVBF (red) under different simulation conditions. (a) $\sigma_v^2 = 0.1, \lambda = 0.95, K = 400$. (b) $\sigma_v^2 = 10, \lambda = 0.95, K = 5000$. (c) $\sigma_v^2 = 10, \lambda = 0.95, K = 5000$ (same as (b) but zoomed out).

to the filtering case

$$\boldsymbol{\Gamma}_k = \mathbb{E}_{\tilde{p}(\mathcal{X}_k)}[\boldsymbol{x}_k \boldsymbol{x}_{k-1}^T], \tag{5.31}$$

$$\boldsymbol{\Lambda}_{k-1} = \mathbb{E}_{\tilde{p}(\mathcal{X}_k)}[\boldsymbol{x}_{k-1} \boldsymbol{x}_{k-1}^T]. \tag{5.32}$$

Unlike DFK algorithms, since expectations of $\boldsymbol{x}_{k-1}$ are taken under the variational state distribution $\tilde{p}(\mathcal{X}_k)$ conditioned on data up to time point $k$, one-step smoothing is required at each iteration. Once again this proceeds on the same lines as the offline VB case for only a single time point. Another difference to DKF algorithms is that (5.24) and (5.28) are evidently coupled so that iterations are required for convergence; one or two iterations usually suffice.

**Results:** For this problem the initial state was assumed to be known and both the DKF and the DVBF were initialised with $\hat{\boldsymbol{x}}_{0|0} = \boldsymbol{x}_0^*$ and $\boldsymbol{\Sigma}_{0|0} = 0.0001\boldsymbol{I}$ where $\boldsymbol{x}_0^*$ denotes the true state at $k = 0$. The parameter vector was initialised to $\hat{\boldsymbol{\theta}}_{0|0} = vec(\boldsymbol{I})$ and for the DVBF $\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_0)}[\boldsymbol{A}_0^T \boldsymbol{\Sigma}_w^{-1} \boldsymbol{A}_0] = 20\boldsymbol{I}$. Both filters were run on data sets generated using the above model with varying observation noise $\sigma_v$ and forgetting factor $\lambda$.

The algorithms showed very similar behaviour at low observation noise levels (Figure 5.2a) and when employing the same forgetting factor. This, however, was to be expected since the crucial difference between the two filters is the utilisation of state and parameter uncertainties in the alternating updates with the DVBF. At these low noise levels it will be seen that the DVBF, on average, still performs marginally better and even with the use of only 1 VB iteration. However, as a result of the one-step smoothing and the increase in matrix computations required, it was noted that the computational time required for the DKF was up to an order of magnitude less than that for the DVBF algorithm. Where noise levels are low, the DKF approach remains the method of choice.

The respective behaviours of the two filters began to diversify in situations of large observation noise, where the state and parameter uncertainties have their effect on estimation. From Figure 5.2b it is seen how the DVBF exhibits an extra degree of caution. This results in a slower convergence time, but a much better overall approximation as seen from Figure 5.2c. It has to be emphasised that the behaviour shown is for the DKF and DVBF methods using the *same forgetting factor*, the added caution in approximation was employed by the DVBF solely as a result of the increased state uncertainty. In this context the DVBF is seen to automatically adjust its learning rate with the observation noise, a feature which is startlingly apparent and which may prove convenient in several online applications.

To validate the improvement in performance, 50 MC runs with varying observation noise, forgetting factor and number of VB iterations were carried out. As a performance index the mean square error (MSE)

$$\text{MSE} = \frac{1}{K - k_1 + 1} \sum_{k=k_1}^{K} \left[ ||\hat{\boldsymbol{x}}_{k|k} - \boldsymbol{x}_k^*||^2 + ||\hat{\boldsymbol{\theta}}_{k|k} - \boldsymbol{\theta}^*||^2 \right], \tag{5.33}$$

was used where $\boldsymbol{x}_k^*$ and $\boldsymbol{\theta}^*$ are the true state vector at time $k$ and the true parameter vector respectively. The data length $K = 5000$ and the start point was set to $k_1 = 2000$ to omit the effects of initial conditions. As seen from Figure 5.3 the MC runs confirmed the increase in accuracy of the DVBF filter in noisy conditions (corresponding tabular values are given in Table 5.1). From the figure it is also evident how the behaviour of the DVBF is much less sensitive to forgetting factors than DKF methods under these conditions. In all cases, except for the case $\sigma_v^2 = 0.1, \lambda = 0.99$, a 2-sample t-test confirmed that the decrease in MSE with the DVBF did not happen by chance ($p < 0.01$). In all cases there

Table 5.1: MSE over 50 MC runs for the DKF and DVBF under different simulation conditions. The confidence intervals shown are at the two-sigma level.

| | $\sigma_v^2 = 0.1$ | |
|---|---|---|
| $\lambda$ | DKF | DVBF |
| 0.95 | 0.077 ± 0.011 | 0.076 ± 0.011 |
| 0.97 | 0.047 ± 0.008 | 0.046 ± 0.008 |
| 0.99 | 0.017 ± 0.004 | 0.017 ± 0.004 |

| | $\sigma_v^2 = 1$ | |
|---|---|---|
| $\lambda$ | DKF | DVBF |
| 0.95 | 0.105 ± 0.014 | 0.094 ± 0.013 |
| 0.97 | 0.071 ± 0.010 | 0.065 ± 0.009 |
| 0.99 | 0.038 ± 0.005 | 0.036 ± 0.005 |

| | $\sigma_v^2 = 10$ | |
|---|---|---|
| $\lambda$ | DKF | DVBF |
| 0.95 | 0.270 ± 0.036 | 0.168 ± 0.018 |
| 0.97 | 0.205 ± 0.027 | 0.144 ± 0.014 |
| 0.99 | 0.142 ± 0.016 | 0.122 ± 0.008 |

was insignificant difference between the results computed with 1, 3 or 5 VB iterations; all results quoted are for a single VB iteration.

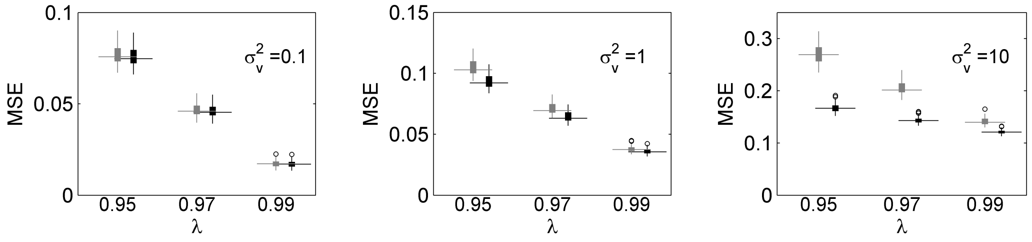
It should be emphasised that alternating standard Kalman filters consitute the *optimal DKF method* to this problem; employing extended, unscented or ensemble methods would not contribute to any increased accuracy in estimation. In this context this example has served as a comparison test between the DVBF and the whole set of possible DKF implementations.

### 5.3.2 Case study: the stochastic diffusion equation

This section demonstrates the potential of the DVBF as an online estimator for varying spatiotemporal dynamics, represented by parameters which are varying both in space *and* in time. It considers once again the diffusion equation of Section 3.3,

$$\frac{\partial z(s,t)}{\partial t} = \frac{\partial}{\partial s}\left(D(s)\left(\frac{\partial}{\partial s} z(s,t)\right)\right) + \sigma \dot{W}(s,t). \tag{5.34}$$

In Section 3.3 the spatially varying parameter $D(s)$ was estimated using batch methods. These methods implicitly assume that the parameters are constant throughout the learning data; parameter changes are not catered for and are, moreover, likely to upset the estimation process. Sequential estimators are excellently suited when there is a suspicion that the dynamic-governing parameters may change.



(a) (b) (c)

Figure 5.3: Box-and-whisker plots of the MSE over 50 MC runs for the DKF (grey) and DVBF (black) under different simulation conditions (a) $\sigma_v^2 = 0.1$, (b) $\sigma_v^2 = 1$ and (c) $\sigma_v^2 = 10$. On each box, the central mark is the median and the edges of the box mark the first and third quartiles. The whiskers extend to the furthest data points that are within 1.5 times the interquartile range. The outliers are plotted individually as circles. In all cases the mean MSE of the DKF is higher than that of the DVBF.

In Theorem 3.1 it was shown that (5.34) may be reduced to the linear model $\boldsymbol{x}_{k+1} = \boldsymbol{A}(\boldsymbol{\vartheta})\boldsymbol{x}_k + \boldsymbol{w}_k$. With sensors taking continuous readings the observation process is also linear and of the form $\boldsymbol{y}_k = \boldsymbol{C}_k\boldsymbol{x}_k + \boldsymbol{v}_k$ (see Section 3.1.2) so that the stochastic diffusion equation is essentially a linear state-space model with a state transition matrix parameterised to cater for spatially-varying effects. Since the DVBF was shown to outperform the DKF in linear state-space models, it will be employed in this study.

For the sake of the exposition, in addition to the states only the parameters composing $\boldsymbol{A}(\boldsymbol{\vartheta})$ are deemed to be unknown so that $\boldsymbol{\theta} = \boldsymbol{\vartheta}$ and the parameter evolution model is given by $\boldsymbol{\vartheta}_k = \boldsymbol{\vartheta}_{k-1} + \boldsymbol{e}_{k-1}$. Noise parameters are therefore assumed to have been found a priori through offline analysis. The filtered state distribution is identical to that given in the previous case study (5.24). The parameter posterior however differs from that of (5.28) as a result of the tailored parameterisation; by application of Theorem 5.2 the required variational parameter posterior may be shown to be equal to

$$\tilde{p}(\boldsymbol{\vartheta}_k) = \mathcal{N}_{\boldsymbol{\vartheta}}(\hat{\boldsymbol{\vartheta}}_{k|k}, \boldsymbol{\Sigma}^{\boldsymbol{\vartheta}}_{k|k}), \tag{5.35}$$

where

$$\boldsymbol{\Sigma}^{\boldsymbol{\vartheta}}_{k|k} = \big(\sigma_w^{-2}\boldsymbol{\Upsilon}_k + \lambda\boldsymbol{\Sigma}^{\boldsymbol{\vartheta}^{-1}}_{k-1|k-1}\big)^{-1}, \tag{5.36}$$

$$\hat{\boldsymbol{\vartheta}}_{k|k} = \boldsymbol{\Sigma}^{\boldsymbol{\vartheta}}_{k|k}\big(\sigma_w^{-2}\boldsymbol{\upsilon}_k + \lambda\boldsymbol{\Sigma}^{\boldsymbol{\vartheta}^{-1}}_{k-1,k-1}\hat{\boldsymbol{\vartheta}}_{k-1|k-1}\big), \tag{5.37}$$

and where $\boldsymbol{\upsilon}_k$ and $\boldsymbol{\Upsilon}_k$ are similar to those in the offline case (3.39) and (3.41) and given by

$$\boldsymbol{\upsilon}_k = \Delta_t[\mathrm{tr}(\boldsymbol{V}_i^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} (\boldsymbol{\Gamma}_k - \boldsymbol{\Lambda}_{k-1}))]_{i=1}^{d}, \tag{5.38}$$

$$\boldsymbol{\Upsilon}_k = \Delta_t^2[\mathrm{tr}(\boldsymbol{V}_i^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{V}_j \boldsymbol{\Lambda}_{k-1})]_{i,j=1}^{d}. \tag{5.39}$$

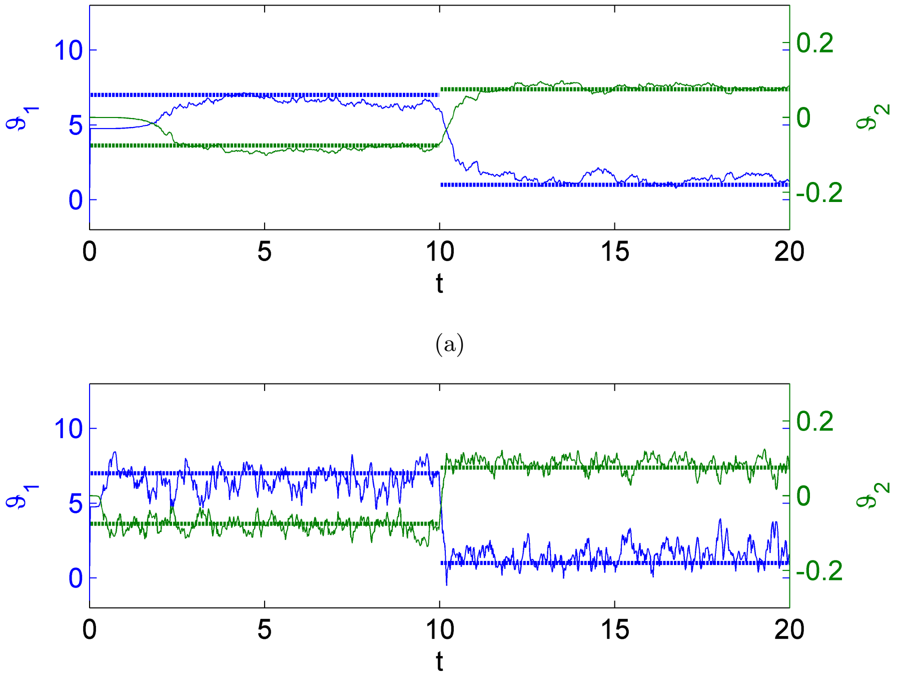
Note how similar the equations (5.36), (5.37) are to the offline case (3.77), (3.78), with the restricted variational posterior $\bar{p}(\boldsymbol{\vartheta}_{k-1})$ taking the role of a prior distribution. The algorithm for online estimation of the dynamics governing the linear SPDEs is given in Algorithm A.4.

Synthetic data for this study was generated on $\mathcal{O} = [0, 60]$ with $n^{sim} = 61$ basis functions, $\tau^{sim} = 4.2$, $\Delta_t^{sim} = 0.02$, $\sigma_w = 0.5$, $\sigma_v = 0.01$ and with $Qu = \int_{\mathcal{O}} k_Q(s - r)u(r)\mathrm{d}r$ with $k_Q(s) = \exp(-s^2/4)$. Unlike in the offline methods, $D(s)$ was not held constant and changed suddenly from $D_1(s) = 7 - 0.1s$ to $D_2(s) = 1 + 0.1s$ at $t = 10$. The sequential estimator was run on data gathered using 51 point sensors with a sample time of $\Delta_t = 0.02$ until $t = 20$, with the role of tracking the spatial variation of the parameters in real-time. The algorithm was initialised with $\hat{\boldsymbol{\theta}}_{0|0} = \hat{\boldsymbol{x}}_{0|0} = \boldsymbol{0}$, $\boldsymbol{\Sigma}_{0|0}^{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{0|0} = \boldsymbol{I}$. It was assumed that a suitable set of basis functions for this system were found a priori.

**Results:** To avoid adverse effects of model mismatch, which were already discussed in Section 3.3.4, at first the same set of basis functions for simulation and estimation was used. As seen from Figure 5.4a and 5.4b in this case the DVBF gave accurate results for forgetting factors between $\lambda = 0.5$ and $\lambda = 0.9$. Whilst higher values for $\lambda$ tended to give too slow a learning rate, smaller values resulted in very noisy estimates. The parameter trajectories initially approximately converge to the true parameters composing $D_1(s)$. As soon as the spatially varying parameter is changed to $D_2(s)$ the parameter trajectories approximately converge to the new true parameter values as desired.

A temporal trajectory of the parameters when using a set of basis functions obtained through frequency analysis, as discussed in Section 3.3.2, is shown in Figure 5.5. As a result of model mismatch the performance is clearly less favourable than when the exact set of basis functions is used, exhibiting considerable fluctuations even with a relatively high forgetting factor of $\lambda = 0.95$. Nonetheless, clear changes in the dynamics are highlighted in the trajectory. The ability to estimate online varying heterogeneous dynamic behaviour of a spatiotemporal system using a model of lower dimensionality than the true model is an encouraging result of this work.

It should be noted that in the absence of any input and dominating initial conditions from which field dynamics become very evident, the online estimation of the spatially varying parameter in noisy conditions is a very challenging task. Unlike in the offline case,



(a)

(b)

Figure 5.4: Sequential estimation of $\vartheta_1$ (blue) and $\vartheta_2$ (green), where $D(s) = \vartheta_1 + \vartheta_2 s$, using the DVBF with the ideal model and (a) $\lambda = 0.9$ and (b) $\lambda = 0.5$. The horizontal dotted lines indicate the true parameter value.

in this scenario the DVBF failed to produce promising results with high observation noise levels, such as $\sigma_v > 0.05$. The limitations will be seen to cause even greater problems with point process observations where a modification to Theorem 5.2 is sought.

## 5.4 Dual VB filtering from temporal point process observations

This work presents for the first time, to the best of the author's knowledge, the application of dual filtering to point process systems. To shed insight on the behaviour on these algorithms with event-based data, the DVBF is therefore first studied with temporal processes. The methods are extended to the spatiotemporal setting in Section 5.5.

### 5.4.1 Homogeneous multi-channel output systems

As in Section 4.1, in this section the events are assumed to be generated by a temporally varying spatially homogeneous intensity which may be represented by a one-dimensional

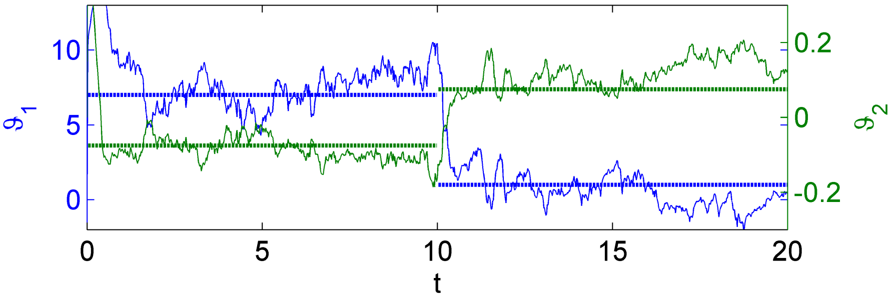


Figure 5.5: Sequential estimation of $\vartheta_1$ (blue) and $\vartheta_2$ (green) using the DVBF with $\lambda = 0.95$ and model constructed using basis functions obtained through frequency analysis as described in Section 3.3.2. The horizontal dotted lines indicate the true parameter value.

state $x_k \in \mathbb{R}$. The state-space dynamic equations are

$$p(x_k \mid x_{k-1}, I_k, \sigma_w^2) = \mathcal{N}_{x_k}(\rho x_{k-1} + \alpha I_k, \sigma_w^2), \tag{5.40}$$

$$p(y_k^c \mid x_k, \mu, \bar{\beta}^c) \approx [\Delta_t \lambda_k^c]^{y_k^c} \exp(-\Delta_t \lambda_k^c), \tag{5.41}$$

where

$$\lambda_k^c = \exp(\mu + \bar{\beta}^c x_k), \tag{5.42}$$

for $c = 1 \dots C$. For this model the set of unknown parameters is given by $\boldsymbol{\theta} = \{(\alpha, \rho), \mu, \bar{\beta}^1, \bar{\beta}^2, \dots, \bar{\beta}^C\}$ the joint of which is assumed to be fully factorised under VB. Since the parameters may be treated separately, the corresponding dynamics are given by

$$\theta_{k+1}^i = \theta_k^i + e_k^i, \tag{5.43}$$

where each

$$e_k^i \sim \mathcal{N}_{e_k^i}\left(0, (\lambda^i)^{-1} \sigma^2_{\theta^i_{k-1|k-1}}\right), \tag{5.44}$$

and $i \in \{\alpha, \rho, \mu, \bar{\beta}^1, \bar{\beta}^2, \dots, \bar{\beta}^C\}$. This formulation allows for a different learning rate for each parameter.

The variational posteriors under RVB are given by

$$\tilde{p}(\mathcal{X}_k) \propto \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta}_k)}[\ln p(\mathcal{X}_k, \mathcal{Y}_k, \boldsymbol{\Theta}_k)]), \tag{5.45}$$

$$\tilde{p}(\alpha_k, \rho_k) \propto \exp(\mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\boldsymbol{\Theta}_k)/(\alpha_k, \rho_k)}[\ln p(\mathcal{X}_k, \mathcal{Y}_k, \boldsymbol{\Theta}_k)]), \tag{5.46}$$

$$\tilde{p}(\mu_k) \propto \exp(\mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\boldsymbol{\Theta}_k)/\mu_k}[\ln p(\mathcal{X}_k, \mathcal{Y}_k, \boldsymbol{\Theta}_k)]), \tag{5.47}$$

$$\tilde{p}(\bar{\beta}_k^c) \propto \exp(\mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\boldsymbol{\Theta}_k)/\bar{\beta}_k^c}[\ln p(\mathcal{X}_k, \mathcal{Y}_k, \boldsymbol{\Theta}_k)]), \qquad c = 1 \dots C, \tag{5.48}$$

which may be computed sequentially using Theorem 5.2. However, for the same reasons discussed in the offline point process case, iterations yield distributions at $k$ which are not of the same form as those at $k-1$ so that further approximations are required to ensure tractability. In the same vein as Section 4.1.2 here it is proposed that Laplace approximations are carried out online (where required) to maintain closed form recursive updates.

**Finding $\tilde{p}(\boldsymbol{x}_k)$:** As in the linear case $\tilde{p}(\boldsymbol{x}_k)$ is updated in the same way as the forward pass in the offline case with VB-Laplace and with expectations taken with respect to $\tilde{p}(\boldsymbol{\theta}_k)$. For technical details refer to Section 4.1.3.

**Online update of $\tilde{p}(\alpha_k)$:** To find the marginal distribution $\tilde{p}(\alpha_k)$, first the joint distribution $\tilde{p}(\rho_k, \alpha_k)$ from (5.46) is written explicitly as

$$\begin{aligned}\tilde{p}(\alpha_k, \rho_k) &\propto \exp(\mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\alpha_{k-1},\rho_{k-1})}[\ln p(x_k|x_{k-1}, \rho_k, \alpha_k)p(\alpha_k|\alpha_{k-1})p(\rho_k|\rho_{k-1})]) \\ &\propto \exp\Bigg( - \mathbb{E}_{\tilde{p}(\alpha_{k-1})}\left[\frac{(\alpha_k-\alpha_{k-1})^2}{2\lambda^{\alpha^{-1}}\sigma^2_{\alpha,k-1|k-1}}\right] - \mathbb{E}_{\tilde{p}(\rho_{k-1})}\left[\frac{(\rho_k-\rho_{k-1})^2}{2\lambda^{\rho^{-1}}\sigma^2_{\rho,k-1|k-1}}\right] \\ &\qquad - \mathbb{E}_{\tilde{p}(\mathcal{X}_k)}\left[\frac{(x_k-\rho_k x_{k-1}-\alpha_k I_k)^2}{2\sigma_w^2}\right]\Bigg).\end{aligned} \tag{5.49}$$

Equation (5.49) may be further re-written as

$$\begin{aligned}\tilde{p}(\alpha_k, \rho_k) \propto \exp\Bigg( &-\frac{\rho_k^2}{2}\left[\frac{1}{\lambda^{\rho^{-1}}\sigma^2_{\rho,k-1|k-1}} + \frac{\widetilde{\lambda}_{k-1}}{\sigma_w^2}\right] + \rho_k\left[\frac{\gamma_k}{\sigma_w^2} + \frac{\hat{\rho}_{k-1|k-1}}{\lambda^{\rho^{-1}}\sigma^2_{\rho,k-1|k-1}} - \frac{\alpha_k g_k}{\sigma_w^2}\right] \\ &-\mathbb{E}_{\tilde{p}(\alpha_{k-1})}\left[\frac{(\alpha_k-\alpha_{k-1})^2}{2\lambda^{\alpha^{-1}}\sigma^2_{\alpha,k-1|k-1}}\right] - \frac{\alpha_k^2 u_k}{2\sigma_w^2} + \frac{\alpha_k r_k}{\sigma_w^2}\Bigg).\end{aligned} \tag{5.50}$$

It helps in the derivation to set

$$\sigma^2_{\rho_k|\alpha_k} = \left[\frac{1}{\lambda^{\rho^{-1}}\sigma^2_{\rho,k-1|k-1}} + \frac{\widetilde{\lambda}_{k-1}}{\sigma_w^2}\right]^{-1}, \tag{5.51}$$

$$\mathbb{E}_{\tilde{p}(\rho_k|\alpha_k)}[\rho_k] = \sigma^2_{\rho_k|\alpha_k}\left[\frac{\gamma_k}{\sigma_w^2} + \frac{\hat{\rho}_{k-1|k-1}}{\lambda^{\rho^{-1}}\sigma^2_{\rho,k-1|k-1}} - \frac{\alpha_k g_k}{\sigma_w^2}\right]. \tag{5.52}$$

The posterior $\tilde{p}(\alpha_k)$ is then found by marginalising $\rho_k$ from the joint to give

$$\begin{aligned}\tilde{p}(\alpha_k) \propto \exp\Bigg( &- \mathbb{E}_{\tilde{p}(\alpha_{k-1})}\left[\frac{(\alpha_k - \alpha_{k-1})^2}{2\lambda^{\alpha^{-1}}\sigma^2_{\alpha,k-1|k-1}}\right] - \frac{\alpha_k^2 u_k}{2\sigma_w^2} + \frac{\alpha_k r_k}{\sigma_w^2} \\ &+ \frac{\sigma^2_{\rho_k|\alpha_k}}{2}\left[\frac{\gamma_k}{\sigma_w^2} + \frac{\hat{\rho}_{k-1|k-1}}{\lambda^{\rho^{-1}}\sigma^2_{\rho,k-1|k-1}} - \frac{\alpha_k g_k}{\sigma_w^2}\right]^2\Bigg), \end{aligned} \tag{5.53}$$

which is normal in $\alpha_k$ with variance and mean

$$\sigma^2_{\alpha,k|k} = \left(\frac{1}{\lambda^{\alpha^{-1}}\sigma^2_{\alpha,k-1|k-1}} + \frac{u_k}{\sigma_w^2} - \frac{\sigma^2_{\rho_k|\alpha_k} g_k^2}{\sigma_w^4}\right)^{-1}, \tag{5.54}$$

$$\hat{\alpha}_{k|k} = \sigma^2_{\alpha,k|k}\left(\frac{\hat{\alpha}_{k-1|k-1}}{\lambda^{\alpha^{-1}}\sigma^2_{\alpha,k-1|k-1}} + \frac{r_k}{\sigma_w^2} - \frac{g_k}{\sigma_w^2}\left[\frac{\gamma_k \sigma^2_{\rho_k|\alpha_k}}{\sigma_w^2} + \frac{\sigma^2_{\rho_k|\alpha_k}\hat{\rho}_{k-1|k-1}}{\lambda^{\rho^{-1}}\sigma^2_{\rho,k-1|k-1}}\right]\right). \tag{5.55}$$

**Finding $\tilde{p}(\rho_k)$:** The statistics over $\rho_k$ are obtained by essentially repeating the procedure described for $\alpha_k$, i.e. first the joint is expressed as

$$\begin{aligned}\tilde{p}(\alpha_k, \rho_k) \propto \exp\Bigg(&-\frac{\alpha_k^2}{2}\left[\frac{1}{\lambda^{\alpha^{-1}}\sigma^2_{\alpha,k-1|k-1}} + \frac{u_k}{\sigma_w^2}\right] + \alpha_k\left[\frac{r_k}{\sigma_w^2} + \frac{\hat{\alpha}_{k-1|k-1}}{\lambda^{\alpha^{-1}}\sigma^2_{\alpha,k-1|k-1}} - \frac{\rho_k g_k}{\sigma_w^2}\right] \\ &-\mathbb{E}_{\tilde{p}(\rho_{k-1})}\left[\frac{(\rho_k - \rho_{k-1})^2}{2\lambda^{\rho^{-1}}\sigma^2_{\rho,k-1|k-1}}\right] - \frac{\rho_k^2\widetilde{\lambda}_{k-1}}{2\sigma_w^2} + \frac{\rho_k\gamma_k}{\sigma_w^2}\Bigg), \end{aligned} \tag{5.56}$$

and then $\alpha_k$ is integrated out to give

$$\begin{aligned}\tilde{p}(\rho_k) \propto \exp\Bigg(&-\mathbb{E}_{\tilde{p}(\rho_{k-1})}\left[\frac{(\rho_k - \rho_{k-1})^2}{2\lambda^{\rho^{-1}}\sigma^2_{\rho,k-1|k-1}}\right] - \frac{\rho_k^2\widetilde{\lambda}_{k-1}}{2\sigma_w^2} + \frac{\rho_k\gamma_k}{\sigma_w^2} \\ &+\left[\frac{1}{\lambda^{\alpha^{-1}}\sigma^2_{\alpha,k-1|k-1}} + \frac{u_k}{\sigma_w^2}\right]^{-1}\left[\frac{r_k}{\sigma_w^2} + \frac{\hat{\alpha}_{k-1|k-1}}{\lambda^{\alpha^{-1}}\sigma^2_{\alpha,k-1|k-1}} - \frac{\rho_k g_k}{\sigma_w^2}\right]^2\Bigg). \end{aligned} \tag{5.57}$$

After some lengthy algebraic manipulations the variance and mean of the required marginal may be found in terms of (5.51), (5.54) and (5.55) as

$$\sigma^2_{\rho,k|k} = \sigma^2_{\rho_k|\alpha_k} + \frac{\sigma^2_{\alpha,k|k}\sigma^4_{\rho_k|\alpha_k} g_k^2}{\sigma_w^4}, \tag{5.58}$$

$$\hat{\rho}_{k|k} = \sigma^2_{\rho_k|\alpha_k}\left[\frac{\gamma_k}{\sigma_w^2} + \frac{\hat{\rho}_{k-1|k-1}}{\lambda^{\rho^{-1}}\sigma^2_{\rho,k-1|k-1}} - \frac{g_k\hat{\alpha}_{k|k}}{\sigma_w^2}\right]. \tag{5.59}$$

Note that Laplace approximations were not required to find $\tilde{p}(\alpha_k)$ and $\tilde{p}(\rho_k)$.

Finally, it is required to find $\mathbb{E}_{\tilde{p}(\alpha_k,\rho_k)}[\rho_k\alpha_k]$ for use with the VB filter. Using (5.50) as a starting point, the joint may be re-written (ignoring unnecessary terms) as

$$\tilde{p}(\alpha_k,\rho_k) \propto -\frac{1}{2}\begin{pmatrix}\rho_k & \alpha_k\end{pmatrix}\begin{pmatrix}\sigma^{-2}_{\rho_k|\alpha_k} & g_k\sigma_w^{-2} \\ g_k\sigma_w^{-2} & \lambda^{\alpha}\sigma^{-2}_{\alpha,k-1|k-1} + u_k\sigma_w^{-2}\end{pmatrix}\begin{pmatrix}\rho_k \\ \alpha_k\end{pmatrix} + \ldots. \tag{5.60}$$

By letting $\lambda^{\alpha}\sigma^{-2}_{\alpha,k-1|k-1}+u_k\sigma_w^{-2} = \sigma^{-2}_{\alpha,k|k}+g_k^2\sigma^2_{\rho_k|\alpha_k}\sigma_w^{-4}$ from (5.54) and taking the inverse of the precision matrix in (5.60), it may be found that $\text{cov}(\rho_k,\alpha_k) = -g_k\sigma_w^{-2}\sigma^2_{\alpha,k|k}\sigma^2_{\rho_k|\alpha_k}$. The required quantity is then readily given (after simplification) as

$$\begin{aligned}\mathbb{E}_{\tilde{p}(\alpha_k,\rho_k)}[\rho_k\alpha_k] &= \text{cov}(\rho_k,\alpha_k) + \hat{\rho}_{k|k}\hat{\alpha}_{k|k} \\ &= \sigma^2_{\rho_k|\alpha_k}[\gamma_k\hat{\alpha}_{k|k}\sigma_w^{-2} - g_k\sigma_w^{-2}(\hat{\alpha}^2_{k|k} + \sigma^2_{\alpha,k|k})].\end{aligned} \tag{5.61}$$

**Finding $\tilde{p}(\mu_k)$:** Following (5.47) and ignoring terms independent of $\mu_k$

$$\begin{aligned}\ln\tilde{p}(\mu_k) &= \mathbb{E}_{\tilde{p}(\mu_{k-1})}[\ln p(\mu_k|\mu_{k-1})] + \mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\bar{\boldsymbol{\beta}}_k)}[\ln p(\boldsymbol{y}_k|x_k,\mu_k,\bar{\boldsymbol{\beta}}_k)], \qquad (5.62) \\ &= -\frac{\mathbb{E}_{\tilde{p}(\mu_{k-1})}[(\mu_k-\mu_{k-1})^2]}{2\lambda^{\mu^{-1}}\sigma^2_{\mu,k-1|k-1}} + \mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\bar{\boldsymbol{\beta}}_k)}\left[\sum_{c=1}^{C} y_k^c[\mu_k+\bar{\beta}_k^c x_k] - \exp(\mu_k)\exp(\bar{\beta}_k^c x_k)\Delta_t\right],\end{aligned}$$

where the state evolution distribution is omitted since it is independent of $\mu_k$. On expanding and approximating around $\hat{\mu}_{k|k}$, the following update equations are obtained

$$\hat{\mu}_{k|k} = \hat{\mu}_{k-1|k-1} + \lambda^{\mu^{-1}}\sigma^2_{\mu,k-1,k-1}\sum_{c=1}^{C}\left(y_k^c - \Delta_t\mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\bar{\boldsymbol{\beta}}_k)}[\exp(\hat{\mu}_{k|k}+\bar{\beta}_k^c x_k)]\right), \tag{5.63}$$

$$\sigma^2_{\mu,k|k} = \left(\lambda^{\mu}\sigma^{-2}_{\mu,k-1|k-1} + \Delta_t\exp(\hat{\mu}_k)\sum_{c=1}^{C}\mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\bar{\boldsymbol{\beta}}_k)}[\exp(\bar{\beta}_k^c x_k)]\right)^{-1}. \tag{5.64}$$

**Finding $\tilde{p}(\bar{\beta}_k^c)$:** Following the same reasoning as that for updating $\tilde{p}(\mu_k)$, the resulting equations are obtained from (5.48) as

$$\begin{aligned}\hat{\bar{\beta}}^c_{k|k} = \hat{\bar{\beta}}^c_{k-1|k-1} + \lambda^{\bar{\beta}^{-1}}\sigma^2_{\bar{\beta}^c,k-1|k-1}\Bigg(& y_k^c\mathbb{E}_{\tilde{p}(\mathcal{X}_k)}[x_k] \\ & - \Delta_t\mathbb{E}_{\tilde{p}(\mu_k)}[\exp\mu_k]\frac{d}{d\bar{\beta}_k^c}\ [\mathbb{E}_{\tilde{p}(\mathcal{X}_k)}[\exp x_k\bar{\beta}_k^c]]\big|_{\bar{\beta}_k^c=\hat{\bar{\beta}}_k^c}\Bigg),\end{aligned} \tag{5.65}$$

$$\sigma^2_{\bar{\beta}^c,k|k} = \left(\lambda^{\bar{\beta}^c}\sigma^{-2}_{\bar{\beta}^c,k-1|k-1} + \Delta_t\mathbb{E}_{\tilde{p}(\mu_k)}[\exp\mu_k]\left[\frac{d^2}{d\bar{\beta}_k^{c^2}}\ \mathbb{E}_{\tilde{p}(\mathcal{X}_k)}[\exp x_k\bar{\beta}_k^c]\big|_{\bar{\beta}_k^c=\hat{\bar{\beta}}_k^c}\right]\right)^{-1}. \tag{5.66}$$

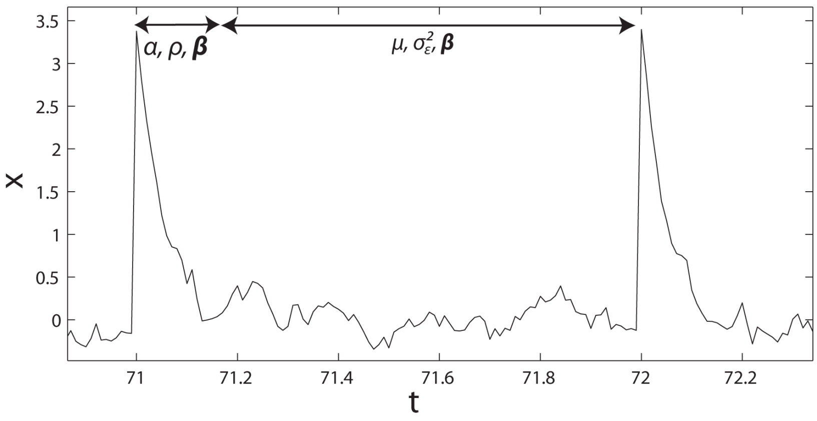


Figure 5.6: Selective updating of parameter estimates in an online framework may be carried out in accordance to the areas where the state bears most information about the relevant parameters of interest. In this case, the narrow stretch close to an input bears a lot of information on the state decay factor $\rho$ and the input gain $\alpha$. The background noise $\mu$ on the other hand, is more evident in regions of no input.

### 5.4.2 Selective estimation

In typical point process systems direct implementation of the above algorithms are likely to be highly inefficient. Take for instance the decay rate of the state $\rho$; in regions of sparse data its estimation is highly uncertain as a direct consequence of the state being unobservable in regions of low event count (see Section 4.1.3, pg. 98). The converse is true for $\mu$; in the presence of a dominant state generating the events (for instance directly after an input), there is little or no information on the background intensity. The result of direct implementation of the DVBF is highly noisy parameter estimations which may be overcome by either i) using very low learning rates (high $\lambda$) or ii) by only selectively updating the parameters in appropriate regions as illustrated in Figure 5.6. The latter procedure has been used previously in online filtering, e.g. speech enhancement by spectral subtraction, in which noise levels are estimated in regions of the signal where speech is not present [219]. Specifying regions for estimation is seen to allow for high learning rates and increase significantly the speed-up for sequential estimation with little or no loss in accuracy.

To demonstrate the effect of the modification consider the point process formulation of (5.40)-(5.43) with $\Delta_t = 0.01$, $K = 20,000$, $C = 20$, $\rho = 0.8$, $\alpha = 4$, $\mu = 0$, with each $\bar{\beta}^c$ a randomly generated number in the interval [0.9 1.1] and an applied spike input every 100 samples. Assume further that $\bar{\boldsymbol{\beta}}$ and $\mu$ were determined from previous analysis and that the algorithm is initialised with $\hat{x}_{0|0} = 0, \sigma^2_{0|0} = 1, \hat{\rho}_{0|0} = 0.6, \sigma^2_{\rho,0|0} = 0.001, \hat{\alpha}_{0|0} = 4$ and

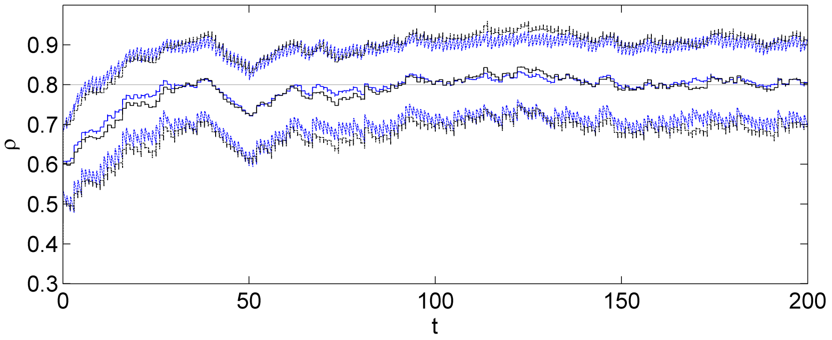


Figure 5.7: Sequential estimation of $\rho$ using a direct implementation of the DVBF (blue) and a modified version which takes into account different regions of operation in the point process system

$\sigma^2_{\alpha,0|0} = 0.01$. The filter results for $\rho$ with $\lambda^\rho = 0.99$ and $\lambda^\alpha = 0.995$ without specifying regions of estimation are shown in Figure 5.7 (blue curve). On the same figure (in black) estimation using only 3 sample time points after the input to estimate $\rho$ and $\alpha$ with $\lambda^\rho = 0.8$ and $\lambda^\alpha = 0.9$ is shown. The mean trajectories and the computed uncertainties between the two DVBFs are virtually indistinguishable. Let the $^*$ superscript be used to denote the true values. The computation times[4] on a standard PC and the MSE error given by

$$\text{MSE} = \frac{1}{K - k_1 + 1} \sum_{k=k_1}^{K} \left[ |\hat{x}_{k|k} - x^*_k|^2 + |\hat{\rho}_{k|k} - \rho^*|^2 + |\hat{\alpha}_{k|k} - \alpha^*|^2 \right], \tag{5.67}$$

for the two approaches with $k_1 = 10,000$ ($t = 100$) are given in Table 5.2, showing drastic savings in computation time for similar estimation quality. The slight difference in MSE should be attributed to different learning rates and not to intrinsic performance difference.

Table 5.2: Estimation time required and quality of estimation using the two versions of the DVBF.

| | Computational time | MSE |
|---|---|---|
| No modification | ∼20mins | 0.1322 |
| Selective estimation | ∼1mins | 0.1319 |

[4] The MATLAB function `fzero` was once again used to find $\hat{x}_{k|k}$.

### 5.4.3 Comparison with SMC methods

In this section the DVBF is compared to a standard PF which makes use of an augmented state vector $\boldsymbol{z}_k = [x_k, \rho_k, \alpha_k]^T$ and implements what is effectively a standard SIS algorithm with resampling (SISR) [see 121, 210, and Section 2.3.1]. The prior distribution is chosen as the importance distribution so that the weights are updated in time according to the likelihood. That is, if $w_k^{(i)}$ denotes the weight of the $i^{th}$ particle at time $k$, and $\boldsymbol{z}_k^{(i)}$ the $i^{th}$ particle at time $k$, the weight update is given as

$$w_k^{(i)} \propto w_{k-1}^{(i)} p(\boldsymbol{y}_k | \boldsymbol{z}_k^{(i)}). \tag{5.68}$$

Resampling is then carried out at each time point resulting in what is known as the *bootstrap filter* [220].

The selective estimation process described above may be adapted to the PF by using *selective* SISR, much in the same vein as employed with the DVBF. In regions where $\rho_k$ and $\alpha_k$ do not affect the likelihood (or importance factor), propagation and subsequent resampling is set to only take place in the state space. The respective parameter posterior distribution is retained and propagated through time unchanged. Formally, after resampling, in this region one has that the full joint distribution is given by

$$p(\alpha_k, \rho_k, x_k | \mathcal{Y}_k) \approx \frac{1}{N} \sum_{i=1}^{N} \delta \begin{pmatrix} x_k - x_k^{(i)} \\ \alpha_k - \alpha_{k-1}^{(i)} \\ \rho_k - \rho_{k-1}^{(i)} \end{pmatrix}, \tag{5.69}$$

and the subsequent posterior distributions by

$$p(\rho_k | \mathcal{Y}_k) = \int p(\alpha_k, \rho_k, x_k | \mathcal{Y}_k) \mathrm{d}x_k \mathrm{d}\alpha_k \approx \frac{1}{N} \sum_{i=1}^{N} \delta(\rho_k - \rho_{k-1}^{(i)}) \approx p(\rho_{k-1} | \mathcal{Y}_{k-1}), \tag{5.70}$$

$$p(\alpha_k | \mathcal{Y}_k) = \int p(\alpha_k, \rho_k, x_k | \mathcal{Y}_k) \mathrm{d}x_k \mathrm{d}\rho_k \approx \frac{1}{N} \sum_{i=1}^{N} \delta(\alpha_k - \alpha_{k-1}^{(i)}) \approx p(\alpha_{k-1} | \mathcal{Y}_{k-1}), \tag{5.71}$$

where $N$ denotes the number of particles and $\delta(\cdot)$ the delta Dirac mass. The selective estimation process is shown in Figure 5.8 where, for illustration, only the input gain $\alpha_k$ and the state $x_k$ are shown. In regions where $\alpha_k$ does not affect the likelihood, propagation and subsequent resampling only takes place in the state-space. The selective resampling procedure results in considerable computational savings for the PF.

Consider a simulation under the same conditions laid out in Section 5.4.2, however with $K = 100,000$ and with $\rho$ changing suddenly from 0.8 to 0.6 at $k = 50,000$. The

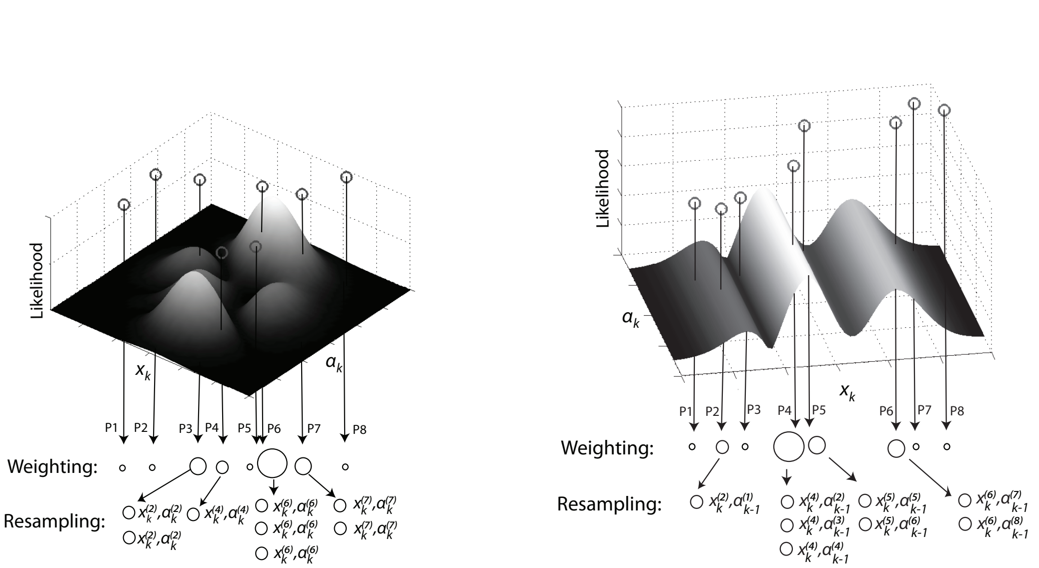


Figure 5.8: (a) The likelihood function is used to appropriately weight the particles (P#) representing the posterior distribution which are then resampled into $N$ particles of equal weight. (b) In this case the likelihood is practically independent of $\alpha_k$ and thus the weighing and resampling steps solely depend on the $x_k$ component of the particles. In order to maintain the posterior distribution with fewer particles than would be necessary otherwise, after resampling, the prior particle parameter set is redistributed with equal weight among the resampled particles. The figures (a) and (b) correspond to the two areas marked in Figure 5.6 respectively (likelihood surfaces shown are for illustration only and do not represent actual surfaces).

forgetting factors of the DVBF were set to $\lambda^\rho = 0.8$ and $\lambda^\alpha = 0.9$. To obtain a similar performance with the PF the factors were set to 0.975 and 0.984 respectively. The PF was configured with $N = 5,000$ particles; the number of particles chosen was the minimum required for consistent posterior distribution approximations across several trials.

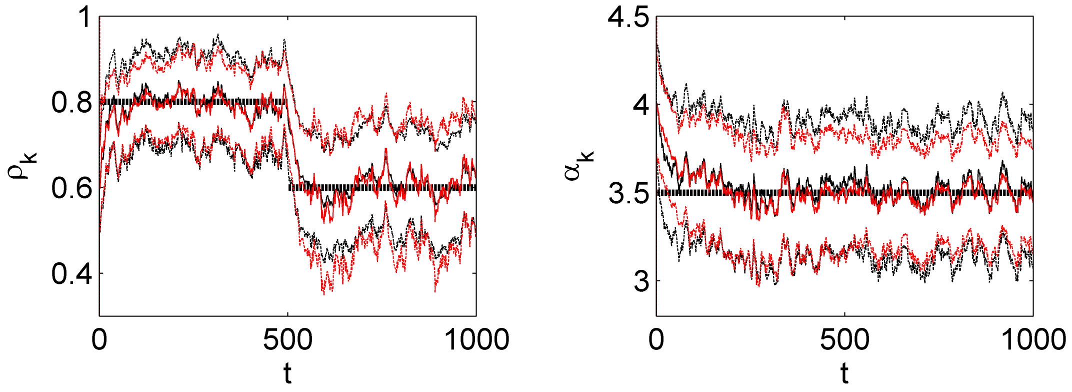
The result for the successful tracking of $\alpha$ and the sudden change in the true value of $\rho$ from 0.8 to 0.6 by both the DVBF and the PF is shown in Figure 5.9. The estimation results are given in Table 5.3 showing similar behaviour for both filters; corresponding computational times and MSEs are given in Table 5.4. In practice, through the use of intermittent resampling (by monitoring the effective sample size) and efficient proposals densities, PFs with better performance may be implemented [121]. However it is considered significant that, despite the parameter distributions estimated being very similar, the PF took an order of magnitude longer than the VB filter to execute.

### 5.4.4 Case study: tastant discrimination from neural responses

As an example application of the novel online algorithm on real data, the spiking patterns of taste-response cells in the nucleus tractus solitarii (NTS) of Sprague-Dawley rats

Table 5.3: Comparison between the VB filter and a PF for SSPP with 5,000 particles.

| | $\rho$ | $\alpha$ | $mean(\hat{\rho}_{k\|k})$ | $mean(\sigma_{\rho,k\|k})$ | $mean(\hat{\alpha}_{k\|k})$ | $mean(\sigma_{\alpha,k\|k})$ |
|---|---|---|---|---|---|---|
| VB (t $\leq$ 500) | 0.8 | 3.5 | 0.799 | 0.037 | 3.52 | 0.13 |
| PF (t $\leq$ 500) | | | 0.797 | 0.031 | 3.49 | 0.12 |
| VB (t $>$ 500) | 0.6 | 3.5 | 0.607 | 0.041 | 3.51 | 0.12 |
| PF (t $>$ 500) | | | 0.602 | 0.049 | 3.49 | 0.10 |



(a)

(b)

Figure 5.9: Online tracking of (a) $\rho$ and (b) $\alpha$ with the true values denoted by the level black lines. In this example $\mu$ and $\beta^c, c = 1 \ldots C$ were assumed constant and known from previous offline analysis of the system. The three-sigma confidence intervals (outer traces) are seen to enclose the true value upon the filter reaching a steady behaviour both for the DVBF with selective estimation (black) and a PF with 5,000 particles (red).

following the application of different taste stimuli [221] are modelled. The attraction of the online approach is that it provides a method for stimulus chemical discrimination by tracking changes in underlying parameters upon the presentation of different stimuli. The experimental data was obtained from trials where different compounds dissolved in distilled water were delivered to the oropharyngeal area. Taste-evoked spike train data used in this study was delivered via neurodatabase.org, a neuroinformatics resource funded by the *Human Brain Project*.

Although the state-space point process model was primarily developed for implicit stimuli, it provides a neat way of parameterising a dynamic CIF to model variable rate neural responses to explicit stimuli. Such is the case considered here, where ample evidence suggests that for some of the cells in the NTS, rate coding is used for inter-

Table 5.4: Estimation time required and quality of estimation using a DVBF and a PF with selective estimation.

| | Computational time | MSE |
|---|---|---|
| DVBF | $\sim$5mins | 0.1220 |
| PF | $\sim$90mins | 0.1221 |

stimulus discrimination [222].[5] Some of these are so fine-tuned to different stimuli that one can use spike count alone to discriminate between different tastes (e.g. cell 9 in the study). Others, on the other hand, are not so fine-tuned and spike count cannot be used to discriminate between the tastants (e.g. cell 11). Nonetheless, spike count gives no information on the time-varying event rate (or rate envelope) itself. Moreover, many alternatives (such as the conventional sliding window) do not provide a plausible model for the underlying neural dynamics. The model applied to these cells not only gives the descriptive powers required for taste discrimination, but also additional information which may be of physiological use. Here it is shown how the DVBF can infer the varying model parameters governing the dynamics, which for the same neuron appear to vary in a structured manner with the application of different stimuli.

Each experimental trial consisted of three phases: i) a 10s baseline period in the absence of any stimulus, ii) 5s of stimulus presentation, and iii) a 5s wait. Each trial was separated by rinsing and a 1.5min wait. The data used in the analysis was that recorded in the second and third phases (10s segments), in which the neural response to the four tastants used, NaCl, HCl, quinine and sucrose, (each of which represents a different taste quality; salty, sour, bitter and sweet respectively), is present. The learning data set was formed by first grouping the 10s segments according to stimulus, and then concatenating them into four sets (1 per stimulus). Combinations of these spike trains were then joined together to form the data sets on which learning was carried out.

Data was gathered at a resolution of 1ms and hence, initially, the spikes were organised into bins of $\Delta_t = 1$ms such that condition (4.1) is satisfied. However the bin size was increased to $\Delta_t = 10$ms to speed up the algorithm. This resulted in some bins ($<$ 5%) containing more than one output spike[6] which were subsequently repositioned to the closest empty bin in forward time. Pre-analysis of the data was carried out by studying the post-stimulus histograms (PSTH) of the responses to the four stimuli. These histograms suggested an approximate linear increase in firing rate for the first 250ms, and

[5] As opposed to temporal coding where the exact time of event incidence is of particular relevance to discrimination.

[6] E.g. for cell 9 - max. HCl with 3.4% and min. sucrose with 1.5%.

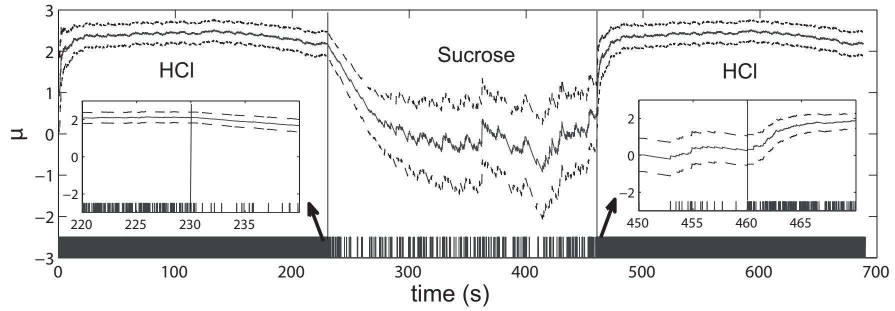


Figure 5.10: Tracking the mean (solid) and corresponding three-sigma intervals (dashed) of $\mu$ indicating a change in stimulus from HCl to sucrose and back to HCl in cell 9. The parameter change is indicative of a change in the spike train pattern (inset) when the stimulus is changed. The solid vertical lines indicate where the change in applied chemical stimulus took place. For this trial $\alpha$ was fixed to 0.1.

also a response latency which was not considered in the simulation study. To cater for these effects, the input signal was treated as a pulse of width 250ms.

It was evident, from preliminary studies, that the dominant rate coding characteristics which differed across tastants were attributed to the input gain $\alpha$ and the background firing rate $\mu$. It was thus deemed appropriate to monitor these two parameters online (in addition to the underlying state) in order to study the response behaviour whilst discriminating between the tastants in real-time. With the use of offline methods the unknown parameters were fixed to $\bar{\beta} = 0.5$, $\sigma^2_\epsilon = 0.05$ and $\rho = 0.97$, which was representative of all tastants. The DVBF was however found to be robust and resistant to changes in state noise and fixed parameter estimates. The relevant forgetting factors were set to $\lambda^\mu = 0.999$ and $\lambda^\alpha = 0.9$ respectively. The initial parameters were set to $\hat{x}_{0|0} = 0, \sigma^2_{0|0} = 1, \hat{\alpha}_{0|0} = 0.1, \sigma^2_{\alpha,0|0} = 0.01, \hat{\mu}_{0|0} = 2.5$ and $\sigma^2_{\mu,0|0} = 0.2$. As discussed in the previous study, the online parameter updates were carried out only in the regions where ample information is present, so that $\alpha$ was only updated in regions of input application and $\mu$ in regions between the application of the respective inputs.

Results from the DVBF show that both the change in $\alpha$ and that in $\mu$ are very evident across the different experiments. In some cases, monitoring $\mu$ is sufficient to characterise the difference in response to different tastants (see Figure 5.10 for a comparison of sucrose with HCl in cell 9). However, this is not the general case, as shown by the trajectories of the mean parameter estimates of $\alpha$ and $\mu$ in Figure 5.11 and 5.12. For instance, whilst $\mu$ seems to vary across tastants in cell 9 (Figure 5.11), the background firing rate in response to NaCl and HCl for cell 11 are fairly similar (Figure 5.12). It is the input gain $\alpha$ which is different between these two responses. By monitoring the parameters $\mu$ and

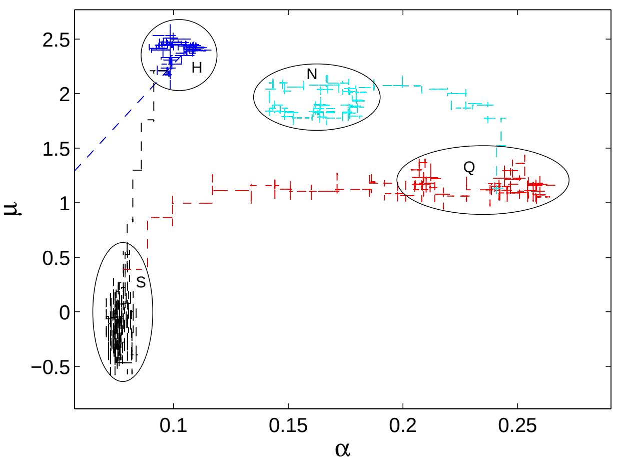


Figure 5.11: Cell 9; temporal progression of the estimated mean of $\alpha$ and $\mu$ indicating a change of stimulus from HCl (H, blue) to sucrose (S, black) to quinine (Q, red) to NaCl (N, cyan). Although the cell is, overall, less responsive ($\mu$) to quinine, the immediate effect of its application ($\alpha$) is relatively more substantial than in the case of both HCl and NaCl. The ellipses define arbitrarily chosen classification boundaries.

$\alpha$, the responses are seen to cluster in distinct and separate regions characteristic to the stimulus being applied.

It is also interesting to note that, except for sucrose, neither response can be considered to be passive (i.e. has both a low $\alpha$ and a low $\mu$). The responses exhibit prominent activity either in the initial stage or the steady-state stage (the phasic and tonic stages respectively), or both. The considerable activity in the initial stage even when the overall response $\mu$ is low (particularly with quinine), is also somewhat of a testimony to the hypothesis that the initial neural response to every tastant may contain some additional information, encoding for instance a measure of taste acceptance (known as the hedonic value, see [221]).

### 5.4.5 Final remarks

In Sections 5.3 and 5.4 the DVBF was shown to outperform conventional Kalman filtering methods, and perform similarly to SMC methods in highly nonlinear processes whilst providing the user with considerable computational savings. It has been applied on real-world data sets where its potential as a tool to learn online parameters in point process systems was noted. The next section will now seek to extend the use of the DVBF to the spatiotemporal point process setting by adding robustness through batch filtering.

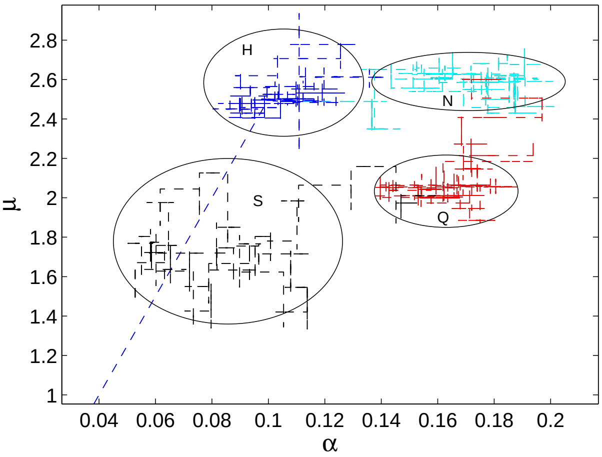


Figure 5.12: Cell 11; temporal progression of the estimated mean of $\alpha$ and $\mu$ indicating a change of stimulus from HCl (H, blue) to NaCl (N, cyan) to quinine (Q, red) to sucrose (S, black). From this chart it is evident that $\alpha$ or $\mu$ on their own cannot capture the difference in response to the different tastants. The ellipses define arbitrarily chosen classification boundaries.

## 5.5 Dual VB filtering from spatiotemporal point process observations

In Section 4.4 it was seen that in the offline case, extracting parameters pertaining to system dynamics from point process observations is a relatively challenging task, typically requiring large data sets for even a few parameters and several iterations for VB convergence. Moreover, in Section 5.4.2 it was implied that in the absence of known inputs (as is commonplace in spatiotemporal systems) the estimation of the dynamics is even harder as the selective estimation as depicted in Figure 5.6 is no longer an option. As a result, direct application of Theorem 5.2 did not work for the spatiotemporal systems studied in Section 4.4, giving parameter estimates highly unrepresentative of the true values even with the wide range of learning rates considered.

The underlying problem lies in the state filtering component of the algorithm. Unlike when continuous observations are available, instantaneous information in the form of events contains negligible information on the state at a specific point in time. Consequently the parameters determine a large part of the a posteriori estimated state (through the predictive distribution), resulting in relatively slow convergence times observed in the offline case. In an online scenario the lack of state information causes irreparable problems: parameters are inferred from highly inaccurate state estimates which in turn, especially in low activity regions, are highly dominated by the previous (inaccurate) parameter estimates.

The problem may be remedied in part by the inclusion of highly informative priors in the online estimation framework. However, a better alternative would be to incorporate more information regarding the state when computing parameters online in spatiotemporal point process systems by considering data in blocks of size $L$. In the EM literature this is possible using a split-data likelihood approach [213]. It is highly desirable, in the context of this work, to derive a filter based on the same concept by establishing a block estimator using variational theory. In the following section it is seen how this entails a simple alteration of Theorem 5.2; the resulting algorithm is the same DVBF which may be easily extended to scenarios where block estimation is required.

### 5.5.1 Block estimation with the DVBF

To maintain the dual estimation framework considered in this chapter, it is required to assume a different, slower, time scale for parameter evolution. This assumption is mild under the premises that the underlying dynamics in spatiotemporal systems vary gradually. Indeed, it is common with spatiotemporal point processes to assume that the parameters do not change at all when inferring the hidden field [35, 37]; the adverse implications of this in outbreak detection are shown in [43]. The consideration of different state/parameter time-scales (slowly time-varying systems) is not new, and has been considered in other application areas such as audio signal enhancement [223].

Consider a temporal index for which $k = rL$ where $L$ is the data block size under consideration. Then for each $r \in \mathbb{Z}^+$, the index $k \in \mathbb{Z}^+$ takes values in $\{(r-1)L\Delta_t + 1, \dots, rL\Delta_t\}$ of length $L$. Now, let $\boldsymbol{\theta}_r := \boldsymbol{\theta}(r\Delta_t)$ and $\boldsymbol{x}_k := \boldsymbol{x}(k\Delta_t)$. Then the state-space dynamic equations are given by

$$\boldsymbol{x}_k = F(\boldsymbol{\theta}_r, \boldsymbol{x}_{k-1}, \boldsymbol{w}_k), \qquad r = \mathrm{int}[(k-1)/L] + 1, \tag{5.72}$$

$$\boldsymbol{y}_k = G(\boldsymbol{\theta}_r, \boldsymbol{x}_k, \boldsymbol{v}_k), \tag{5.73}$$

where int[·] returns the integer part of the argument. The resulting graphical representation of the model is given in Figure 5.13 where the observations are omitted for sake of clarity. The parameter evolution model remains unchanged:

$$\boldsymbol{\theta}_r = \boldsymbol{\theta}_{r-1} + \boldsymbol{e}_{r-1}. \tag{5.74}$$

Let the parameter set $\boldsymbol{\Theta}_r = \{\boldsymbol{\theta}_i\}_{i=1}^{r}$; then the resulting variational approximation to

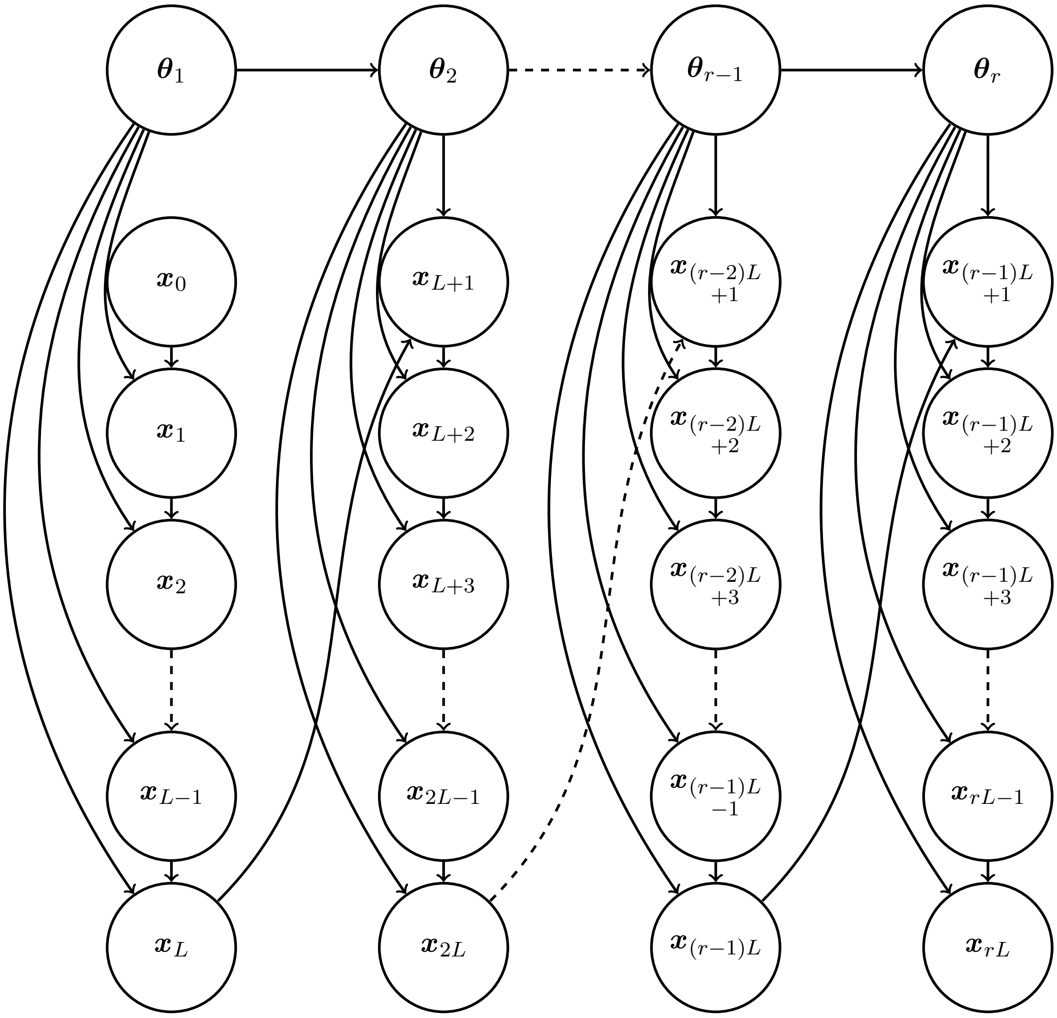


Figure 5.13: Graphical representation of the parameter and state-space latent process under consideration for the block DVBF. Each column behaves as a separate entity to be processed using conventional batch estimation techniques, with prior distributions following from the statistics computed from the previous block. Inference on each block is carried out only when all the relevant data becomes available, i.e. when $k$ is an exact multiple of $L$.

the true posterior distribution $p(\mathcal{X}_k, \boldsymbol{\Theta}_r|\mathcal{Y}_k)$ is given by

$$\tilde{p}(\mathcal{X}_k, \boldsymbol{\Theta}_r) \approx \tilde{p}(\mathcal{X}_k)\prod_{j=1}^{r}\tilde{p}(\boldsymbol{\theta}_j) = \tilde{p}(\mathcal{X}_k)\tilde{p}(\boldsymbol{\Theta}_r). \tag{5.75}$$

Assume further that the parameter $\boldsymbol{\theta}_r = \{\theta_r^1, \theta_r^2, \ldots, \theta_r^l\}$. The optimal variational posteriors are given by

$$\tilde{p}(\mathcal{X}_k) \propto \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta}_r)}[\ln p(\mathcal{X}_k, \mathcal{Y}_k, \boldsymbol{\Theta}_r)]), \tag{5.76}$$

$$\tilde{p}(\theta_r^i) \propto \exp(\mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\boldsymbol{\Theta}_r^{/\theta_r^i})}[\ln p(\mathcal{X}_k, \mathcal{Y}_k, \boldsymbol{\Theta}_r)]), \qquad i = 1 \ldots l. \tag{5.77}$$

A recursive solution for (5.76) and (5.77) is not possible without further modification. In the same way as Section 5.2.1, one may apply RVB to the parameter variational distributions. Rendering explicit the subset of the data with which the variational posterior was computed $\tilde{p}(\boldsymbol{\Theta}_r)$ is now given as

$$\begin{aligned}\tilde{p}(\boldsymbol{\Theta}_r) &= \tilde{p}(\boldsymbol{\theta}_r|\mathcal{Y}_k)\prod_{j=1}^{r-1}\bar{p}(\boldsymbol{\theta}_j|\mathcal{Y}_{jL}) \\ &= \tilde{p}(\boldsymbol{\theta}_r|\mathcal{Y}_k)\bar{p}(\boldsymbol{\Theta}_{r-1}).\end{aligned} \tag{5.78}$$

Once again, as a result of the parameter factorisation, the restricted variational posteriors are given by

$$\bar{p}(\boldsymbol{\Theta}_{r-1}) = \prod_{j=1}^{r-1}\prod_{i=1}^{l}\bar{p}(\theta_j^i|\mathcal{Y}_{jL}). \tag{5.79}$$

The key alteration to Theorem 5.2 lies in the compound estimation of a block of states, rather than a single state at time $k$. Whilst in the simple sequential implementation a recursion is found between $\tilde{p}(\boldsymbol{x}_k)$ and the components of $\tilde{p}(\boldsymbol{\theta}_k)$, in the block case a recursion between $\tilde{p}(\boldsymbol{x}_{(k-L+1):kL})$ and the components of $\tilde{p}(\boldsymbol{\theta}_r)$ is envisaged. The theorem is as follows:

**Theorem 5.3** *For the state-space equations (5.72) and (5.73), given the factorisation (5.75), the restriction (5.79) and the maximisers (5.76) and (5.77), the recursive updates for the state and parameter variational distributions $\tilde{p}(\mathcal{X}_k)$ and $\tilde{p}(\theta_k^i), i = 1 \ldots l$, in blocks of length $L$ are given by*

$$\tilde{p}(\boldsymbol{x}_{k-L+1:k}) \propto p^*(\boldsymbol{x}_{k-L+1})\exp\left(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_r)}\left[\ln\prod_{i=k-L+1}^{k}p(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_r)\prod_{i=k-L+2}^{k}p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1},\boldsymbol{\theta}_r)\right]\right), \tag{5.80}$$

$$\begin{aligned}\tilde{p}(\theta_r^j) \propto \exp&\left(\mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\boldsymbol{\theta}_r^{/j})}\left[\ln\prod_{i=k-L+1}^{k}p(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_r)p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1},\boldsymbol{\theta}_r)\right]\right) \\ &\times\ \exp\left(\mathbb{E}_{\bar{p}(\theta_{r-1}^j)}[\ln p(\theta_r^j|\theta_{r-1}^j)]\right), \qquad j = 1 \ldots l.\end{aligned} \tag{5.81}$$

*where*

$$p^*(\boldsymbol{x}_{k-L+1}) = \int \tilde{p}(\boldsymbol{x}_{k-L}) \exp\left(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_r)}[\ln p(\boldsymbol{x}_{k-L+1}|\boldsymbol{x}_{k-L}, \boldsymbol{\theta}_r)]\right) d\boldsymbol{x}_{k-L}, \tag{5.82}$$

*is the predictive distribution of* $\boldsymbol{x}_{k-L+1}$ *under* $\boldsymbol{\theta}_r$*.*

*Proof.* By marginalisation of the states associated with previous data blocks, one obtains

$$\begin{aligned}\tilde{p}(\boldsymbol{x}_{k-L+1:k}) &\propto \int \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta}_r)}[\ln p(\mathcal{X}_k, \boldsymbol{\Theta}_r, \mathcal{Y}_k)])\mathrm{d}\mathcal{X}_{k-L} \\ &= \exp\left(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_r)}\left[\ln \prod_{i=k-L+1}^{k} p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_r) \prod_{i=k-L+2}^{k} p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}, \boldsymbol{\theta}_r)\right]\right) \\ &\quad \times \underline{\int \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta}_r)}[\ln p(\boldsymbol{x}_{k-L+1}|\boldsymbol{x}_{k-L}, \boldsymbol{\theta}_r) p(\mathcal{X}_{k-L}, \boldsymbol{\Theta}_r, \mathcal{Y}_{k-L})])\mathrm{d}\mathcal{X}_{k-L}}.\end{aligned} \tag{5.83}$$

By comparing to (5.80) it is required to show that the underlined quantity in (5.83) is equal (up to a normalisation constant) to $p^*(\boldsymbol{x}_{k-L+1})$. To this end the quantity is re-expressed as

$$\begin{aligned}&\int \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta}_r)}[\ln p(\boldsymbol{x}_{k-L+1}|\boldsymbol{x}_{k-L}, \boldsymbol{\theta}_r) p(\mathcal{X}_{k-L}, \boldsymbol{\Theta}_r, \mathcal{Y}_{k-L})])\mathrm{d}\mathcal{X}_{k-L} \\ &\propto \int \Bigg[ \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_r)}[\ln p(\boldsymbol{x}_{k-L+1}|\boldsymbol{x}_{k-L}, \boldsymbol{\theta}_r)]) \\ &\qquad \times \int \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta}_{r-1})}[\ln p(\mathcal{X}_{k-L}, \boldsymbol{\Theta}_{r-1}, \mathcal{Y}_{k-L})])\mathrm{d}\mathcal{X}_{k-L-1}\Bigg] \mathrm{d}\boldsymbol{x}_{k-L} \\ &= \int \Bigg[ \exp\left(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_r)}[\ln p(\boldsymbol{x}_{k-L+1}|\boldsymbol{x}_{k-L}, \boldsymbol{\theta}_r)]\right) \\ &\qquad \times \int \underline{\left(\int \exp\left(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta}_{r-1})}[\ln p(\mathcal{X}_{k-L}, \boldsymbol{\Theta}_{r-1}, \mathcal{Y}_{k-L})]\right) \mathrm{d}\mathcal{X}_{k-2L}\right)} \mathrm{d}\boldsymbol{x}_{k-2L+1:k-L-1}\Bigg] \mathrm{d}\boldsymbol{x}_{k-L}.\end{aligned} \tag{5.84}$$

where, as a result of RVB, $\boldsymbol{\Theta}_{r-1}$ does not need to be re-estimated under the new data. The recursion is now obtained by realising that the underlined quantity in (5.84) is equal to the first line of (5.83) with $r$ and $k$ replaced by $r-1$ and $k-L$ respectively. One then obtains the quantity

$$\begin{aligned}
&\int \left[ \exp\left(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_r)}[\ln p(\boldsymbol{x}_{k-L+1}|\boldsymbol{x}_{k-L}, \boldsymbol{\theta}_r)]\right) \int \tilde{p}(\boldsymbol{x}_{k-2L+1:k-L}) \mathrm{d}\boldsymbol{x}_{k-2L+1:k-L-1} \right] \mathrm{d}\boldsymbol{x}_{k-L} \\
&= \int \left[ \tilde{p}(\boldsymbol{x}_{k-L}) \exp\left(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_r)}[\ln p(\boldsymbol{x}_{k-L+1}|\boldsymbol{x}_{k-L}, \boldsymbol{\theta}_r)]\right) \right] \mathrm{d}\boldsymbol{x}_{k-L} \\
&= p^*(\boldsymbol{x}_{k-L+1}), \qquad (5.85)
\end{aligned}$$

as required.

An expression for the optimal variational parameter posterior is obtained as

$$\begin{aligned}
\tilde{p}(\theta_r^j) &\propto \exp\left( \mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\boldsymbol{\Theta}_r^{/\theta_r^j})} [\ln p(\mathcal{X}_k, \mathcal{Y}_k, \boldsymbol{\Theta}_r)] \right) \\
&= \exp\left( \mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\boldsymbol{\Theta}_r^{/\theta_r^j})} \left[ \ln \prod_{i=k-L+1}^{k} p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_r) p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}, \boldsymbol{\theta}_r) \right] \right) \\
&\quad \times \exp\left( \mathbb{E}_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\boldsymbol{\Theta}_r^{/\theta_r^j})} [\ln p(\boldsymbol{\theta}_r|\boldsymbol{\theta}_{r-1}) p(\mathcal{X}_{k-L}, \boldsymbol{\Theta}_{r-1}, \mathcal{Y}_{k-L})] \right), \qquad (5.86)
\end{aligned}$$

from which (5.81) follows since the distribution of interest is independent of $p(\mathcal{X}_{k-L}, \boldsymbol{\Theta}_{r-1}, \mathcal{Y}_{k-L})$ and since $\boldsymbol{\theta}$ is composed of conditionally independent parameter elements. ■

**Remark 5.2** *It is easily seen that the online block DVBF reduces to the DVBF of Theorem 5.2 by setting $L = 1$ and, correspondingly, $r = k$ in (5.80) and (5.81).*

The significance of the online block estimator lies in the structure of the derived variational posteriors. In particular, the optimal variational posteriors of (5.80) and (5.81) may be recognised as those obtained using the usual batch or offline VB algorithm (see (4.92) and (4.93)) with priors over the states and the parameters propagated from posterior distributions computed in the previous block. Thus the block algorithm reduces to a sequence of offline estimations with the state prior being the predictive distribution of the final state of the previous time block and the parameter prior being the predictive distribution of the evolving parameters. The algorithm allows for user-defined $L$, which would typically be set large in the presence of uninformative data (sparse events).

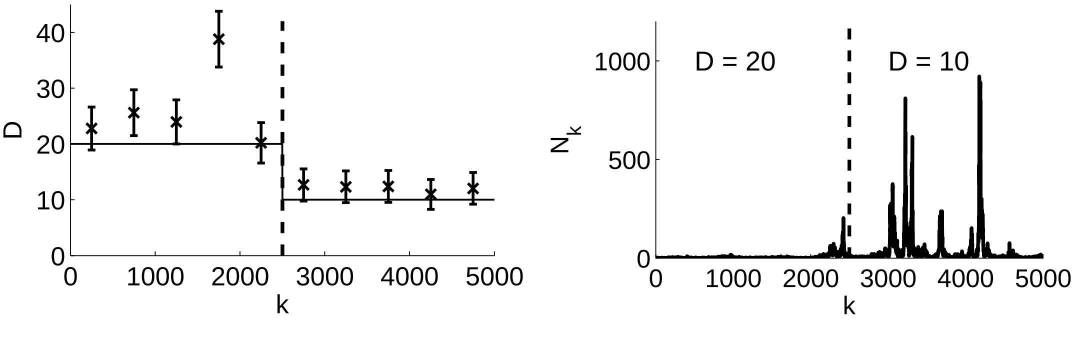
### 5.5.2 Case study: the stochastic heat equation

This section studies the performance of the block DVBF in the context of spatiotemporal point process surveillance. Consider once again the stochastic heat equation with operator $\mathcal{A}(\cdot) = D\Delta(\cdot), D \in \mathbb{R}^+$ studied using offline analysis in Section 4.5. In general, the diffusion constant $D$ is an indication of how quickly spurious hot spots in the latent field disperse and a high $D$ corresponds to lower event intensities under stationarity. In a number of scenarios, such as epidemiology, it may be important to detect when this latent 'energy' is not being dispersed quick enough; an effect which would manifest itself in $D$ becoming smaller and in extreme cases close to zero.

Most other conventional methods attempt to anticipate outbreaks by studying event count or by checking the probability of exceedances (with respect to some threshold) of the inferred latent process [43]. However these approaches do not consider the dynamic nature of the process. More importantly, they do not take into consideration the fact that the dynamics may change *before* the event count or inferred field experiences a noticeable change in magnitude. Because of this, monitoring the parameters which govern the dynamics of spatiotemporal point processes may prove considerably advantageous over conventional approaches.

Let $\mathcal{O} = [0, 30] \times [0, 30] \subset \mathbb{R}^2$. A mesh of $n^{sim} = 49$ basis function of the form (3.20) with $\tau^{sim} = 1.05$ were equally spaced on $\mathcal{O}$ satisfying the Dirichlet boundary conditions. Simulation parameters were set as $\Delta_t = 0.01$, $\sigma_w = 4$, $\mu = -4$, $\bar{\beta} = 1$, $Qu = \int_{\mathcal{O}} k_Q(\boldsymbol{s} - \boldsymbol{r})u(\boldsymbol{r})\mathrm{d}\boldsymbol{r}$ and with $k_Q(\boldsymbol{s}) = \exp(-\boldsymbol{s}^T\boldsymbol{s}/4)$. A data set consisting of $K = 5,000$ times frames was generated with $D = 20$ initially and $D = 10$ from $k = 2,500$ onwards. Once again, simulation of the point process was carried out by the method of thinning and space-time stationarity was assumed. It was also assumed that the set of basis functions and all disturbance parameters were found by employing previous offline analysis (see Chapter 4). The sole aim of the DVBF, in this case, was to monitor the parameter $\vartheta = D$ and to detect when it lowers, preferably before a noticeable change in event count. The DVBF was initialised with $\hat{\vartheta}_{0|0} = 15$ and $\sigma^2_{\vartheta,0|0} = 1000$.

The event rate for the first 2,000 time points was, on average, two events per time frame. Such a low event count necessitated the consideration of large blocks in order to ensure robustness; $L = 500$ was found to be a suitable choice. Note that this number can be drastically reduced in regions of high event count per frame, say 200 counts per frame as in [37]. The forgetting factor $\lambda$ was, in turn, set to 0.1. The low value was chosen in order to offset the slow adaptation which would otherwise be imposed by the large $L$. Combined field-parameter estimation of each block of 500 time points took on the order of 1 day with the DVBF assuming convergence when $|\vartheta^{(i+1)} - \vartheta^{(i)}| < 0.01$. Due to time constraints a direct comparison of this performance with that of the widely used MALA algorithm [185, 35, 37] was not carried out; however the latter has been reported to take 'overnight' [43] for smoothing (with no parameter estimation) of data collected over 5 days (equivalent to a block size of $L = 5$ if $\Delta_t = 1$ day).



(a) (b)

Figure 5.14: (a) Online estimation of the parameter $D = \vartheta$ in blocks of size $L = 500$ using the DVBF. The bars show the three-sigma confidence intervals, the cross the VB posterior mean and the thin line the true parameter trajectory. (b) Number of events at each time point $k$. In both subfigures, the dashed vertical line corresponds to the change-point of the true parameter $D$ from 20 to 10.

**Results:** The trajectory of the true parameter $D_r = \vartheta_r$ and its estimate for this case study are given in Figure 5.14a. As expected there is some variation in the estimation trajectory, especially in regions when $D$ is high and correspondingly the intensity is very low. In particular, the estimation of the $4^{th}$ block is seen to be unrepresentative; the intensity was so low here that virtually nothing could be inferred about the dynamics. In such regions, the parameter will be overestimated (see Footnote 9 on pg. 83) and therefore does not present a problem in this particular scenario. Of more concern is that the credibility intervals do not consistently enclose the true value. This, however, was expected to be the case in regions of low intensity where VB begins to be highly overconfident (see Section 3.4). Of importance to this discussion is that the change in parameter to a *lower* $D$ (corresponding to less energy dissipation) is detected in a timely fashion *before* the escalation in event count (see Figure 5.14b).

The change in dynamics would not have been detected using solely event count. Consider Figure 5.15 which focuses on the time interval where the transition took place. The number of events between $k = 2,000$ and $k = 2,500$ ($5^{th}$ block) was 9,325, nearly double the count between $k = 2,500$ and $k = 3,000$ ($6^{th}$ block) which totalled only 4,746. Based solely on event count, natural conclusions are that a change point occurred at the beginning of the $5^{th}$ block and that in the $6^{th}$ block the system was slowly restoring itself to normality. These conclusions are evidently erroneous since $D$ decreased at the beginning of the $6^{th}$ block.

Methods using probabilities of exceedance would also not have worked in this particular scenario since the maximum posterior mean of the latent field in the $5^{th}$ block was larger than that in the $6^{th}$ block for 65% of the time points. Thus any threshold

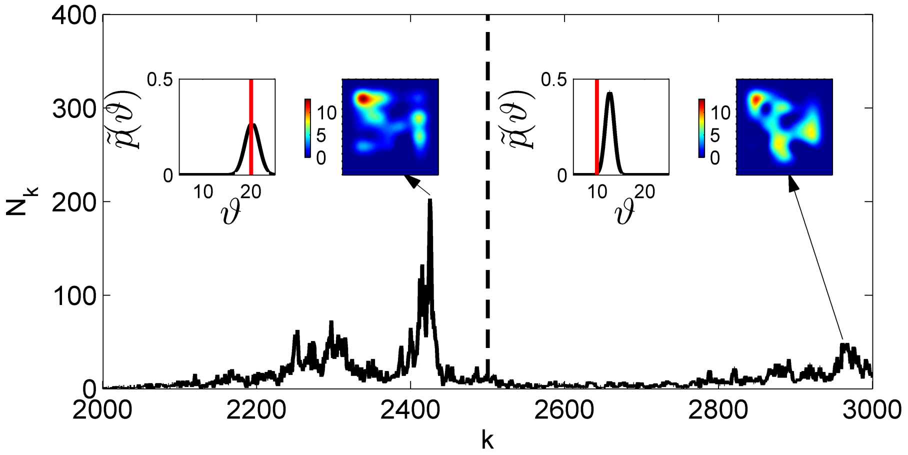


Figure 5.15: Number of events $N_k$ at each time point $k$ close to the transition period from $D = 20$ to $D = 10$ at $k = 2,500$. Insets: The line plots show the variational posteriors $\tilde{p}(\vartheta)$ for the $5^{th}$ and $6^{th}$ estimation blocks (corresponding to $k = 2,000$ to $k = 2,500$ and $k = 2,500$ to $k = 3,000$ respectively) together with the true parameter value (red line). The surface plots show the a posteriori underlying field $z_k$ at the peaks of the two blocks.

employed would likely have been exceeded (in probability) in the $5^{th}$ block more often than in the $6^{th}$ block; again resulting in conclusions converse to what is required.

The advantage of employing a dynamic model and estimating $D$ online in such systems is thus apparent; it may not be the number of events observed which is of concern, but the spatiotemporal evolution of their behaviour. In this case the DVBF successfully detected the shift in dynamics at the $6^{th}$ block. As seen from the event count progression in Figure 5.14b this detection was a real forewarning of things to come.

## 5.6 Conclusion

This chapter has introduced the concept of dual filtering in the context of spatiotemporal systems. This has resulted in several contributions, namely

- the novel concept of online field-parameter estimation in both temporal and spatiotemporal point process systems,
- the development of a new dual variational filter which compares favourably to conventional methods such as DKF methods and SMC methods and
- the extension of the dual variational filter to a block filter, which is theoretically shown to constitute a concatenation of sequential batch estimators.

The developed DVBF was seen to perform particularly well in SPDEs with continuous observations, less so in point process systems when the number of events is relatively low.

In the latter case variability in parameter estimates was quite large and necessitated the use of block filtering. Fortunately, by showing that the DVBF is a special case of the block filter, its implementation does not require the need for significant theoretical and algorithmic extensions.

The viability of the proposed filter for spatiotemporal systems was made possible through the continuous-space framework stablished in Chapter 4 which decouples the discretisation needed for numerical integration from the resolution at which points are being observed. It should be pointed out that adopting a discrete-space representation (with the restriction of a maximum of one event per grid location) for online inference is even less desirable than with offline systems where computational speed is more of an issue and where the data is not available before estimation.

The work carried out in this chapter is a natural extension to the concepts introduced in Chapter 4, where the estimation of the dynamics of spatiotemporal point processes was carried out offline. Since this is the first work of its kind, only simple systems were analysed in these two chapters; these nonetheless serve as a proof of principle for more large-scale systems such as that treated in the next chapter.

There are several paths for future work. First, a strategy for choosing the block size $L$ may be designed, which favours a large $L$ in regions of low event count. For this several studies will be required to quantify the effect of $L$ on the inference quality. Second, given the difficulty encountered in parameter estimation, it might be worthwhile considering an interacting multiple model (IMM) approach [224] in order to select the most likely model from a finite number of candidates. Lastly, through the representation of spatiotemporal systems as a dynamic model, it is now possible to investigate their control, obviously in an online context, using the variational filter or even the DVBF. This would be particularly interesting in the context of point process systems, where the idea of control, to the best of the author's knowledge, has not yet been introduced.

Chapters 3, 4 and 5 have developed and implemented an approach for spatiotemporal system analysis based on concepts emerging from systems theory. The framework delivers a dynamic model in state-space form learnt offline or online from data which may be continuous or in the form of events. The developed theory will be applied next to a real data set; here the dividends of adopting this systems approach to spatiotemporal analysis will become readily apparent.

# Chapter 6

# Modelling and prediction in conflict: Afghanistan

Chapter 4 introduced a novel framework for the study of spatio-temporal point process systems through the use of finite dimensional reduction methods and advanced signal processing techniques. The methodology was seen to perform well on event-based simulated data with high levels of unpredictability and able to extract patterns which are not immediately apparent. It was also fast and memory efficient and handled large data sets with ease.

This chapter shows the efficacy of the developed method in a real-world application, conflict analysis. In essence, a conflict zone can be viewed as a highly unpredictable system with several random (unknown) external influences. All reported incidents in the scene of conflict are usually presented to analysts in the form of war logs so that the data sets under consideration are effectively point processes. Moreover the number of reported incidents usually reaches into the several thousand so that a flexible and computationally efficient method for highlighting emerging trends is highly desirable. In the light of this, the method developed in Chapter 4 is ideal for the study of such data sets.

The proposed methodology applied to conflict analysis is demonstrated on the Afghan War Diary (AWD), a compendium of military war logs released by the whistleblower site Wikileaks in 2010 pertaining to the war in Afghanistan which commenced in 2001. Studies carried out to date of the AWD fail to provide insights into the *dynamics* of the logged events (e.g. [225]) which may play a vital role in explaining the conflict progression and also in predicting future events. The adopted approach in this chapter remedies this by studying the dynamics of the events, thus allowing for optimal interpolation and statistically sound predictions. The methodology is new to conflict analysis and the promising results are indicative of its potential in treating large event-based data sets with similar properties.

To familiarise the reader with the context of the study and the light in which the results should be interpreted, a brief history of the war to date in Afghanistan is given in Section 6.1.1. Section 6.1.2 describes the role of data analysis in a modern-day conflict scenario, the challenges faced by academics in carrying out their own analysis and the role played by the AWD in this regard. Section 6.1.3 describes the work which has been carried out on the AWD and explains why further work is appropriate. Section 6.2 contains a rigorous statistical analysis to support the hypothesis that the logged events in the AWD may be treated as the realisation of a special kind of stochastic process. The learning of the dynamic model using point process statistics and VB inference is carried out in Section 6.3 which also includes some minor modifications from the treatment in Chapter 4 for this special case. A discussion of the results and a demonstration of the model's predictive capabilities is given in Section 6.4; Section 6.5 concludes the chapter.

## 6.1 War in Afghanistan and the Afghan War Diary

### 6.1.1 Brief history of the conflict in Afghanistan

October 7, 2001, saw the beginning of *Operation Enduring Freedom* in Afghanistan, an initiative of the United States of America (USA) and the United Kingdom to dismantle the Al-Qaeda terrorist organisation and remove the Taliban from power. By December 2001 all major Al-Qaeda and Taliban leaders had either been killed or forced to retreat into remote areas and neighbouring country Pakistan. In a little more than two months the USA had achieved their primary goals in Afghanistan and together with the United Nations helped introduce the Afghan Interim Authority and the International Security Assistance Force (ISAF) with the role to maintain national security.

The two-month successful foray was however only the beginning of a long drawn-out war which has plunged Afghanistan into violence and chaos [226]. The year 2002 saw the Taliban and Al-Qaeda regroup their forces deep in the mountains and across the border in Pakistan. They recruited militants to fight what they called the new *jihad* or holy-war against the USA, ISAF and the Afghan government and trained them in guerrilla warfare. By the summer of 2003 the rejuvenated insurgency was carrying out regular ambushes, attacks and raids, resulting in hundreds of fatalities and thousands of wounded. In 2006 the situation took a drastic turn to the worse, with the violence in the south reaching unprecedented levels. According to *http://www.icasualties.org* 331 coalition military personnel perished between 2001 and 2005, and a staggering 1950 between 2006 and 2010. Civilian fatality figures fare even worse, with an estimate of around 4,000 fatalities between 2004 and 2009 [227].

The situation to date remains dire. In January 2011 the Afghanistan NGO (non-governmental organisation) Safety Office (ANSO)[1] reported that there is ‘indisputable evidence that conditions are deteriorating’ [228] after registering a record average of 33 armed attacks *per day* by armed opposition groups (AOG) in 2010. The attacks aimed at killing government personnel and disrupting supply lines but were reportedly also responsible for 83% of the c. 2,000 civilian fatalities in 2010. ANSO also stated that ‘going in to 2011, the AOG position looks strong’. Indeed, according to *www.icasualties.org* the number of coalition fatalities between January and May 2011 was practically the same for the same period in 2010 (220 in 2010 and 216 in 2011). These impressive figures compel the public and taxpayers to query the ultimate motifs of the war and also beckon the need for an unprejudiced quantitative assessment of future prospects. Unbiased assessments are however only possible with unaltered data, which unfortunately is not usually publicly released.

### 6.1.2 Conflict and the data revolution

There is ample evidence today (see section 6.1.3) that the USA military is in possession of a virtually interminable stream of data to aid the key players and policy makers in decision making. The amount of data available is immense; everything is recorded and logged, from suspicious activity to gunfights lasting several hours. Hidden in these logs are patterns and trends explaining the current state of affairs and containing invaluable information about what may possibly take place in the near future [229, Slide 18, ‘Time is Running Out’]. The military tries to anticipate the actions of the aggressors in order to counter them in the most efficient way possible with minimal loss of life and resources. The Pentagon, for instance, today enjoys a yearly budget of approximately US$28 million for the modelling and prediction of insurgency and other aspects of warfare [230].

The reason why the same sort of investment, or interest, is lacking in the academic research community may be pinned down to the nature of the data which is accessible. The overwhelming majority of data logged by the military is labelled secret or classified and is not available to the public, and academics have to rely on unclassified documents or the media to make their own conclusions on a conflict’s state of affairs [231]. These secondary sources are largely idiosyncratic in nature and may be considered unreliable at best:

[1] ANSO is a non-profit organisation whose sole role is to provide security advice for NGOs operating in Afghanistan by issuing bi-weekly and quarterly reports on aggression incidents in the country.

> “As a result of the uncertain reliability of these sources, policy suggestions and academic analysis are always subject to the criticism that the data from which conclusions are drawn have been falsified or are biased, uneven in coverage, amnesiac about certain subjects, or exculpatory of government decisions.” [225].

Consider the present situation in Afghanistan. In 2011 ANSO noted that the current security situation in Afghanistan is alarming and in sharp contrast with the reassuring military coalition’s claims which are ‘solely intended to influence American and European public opinion’ [228]. The media is also not a reliable source of information; for instance, the news network Al Jazeera reported that the media was asked not to report any violent incidents in the Afghanistan 2009 elections despite several occurrences [232]. Academic researchers thus do not have the budget nor the ‘reams of information’ [230] the Department of Defence has at its disposal to make effective qualitative assessments.

The AWD released by Wikileaks on 25 July 2010 is hence a treasure trove for conflict analysts. It contains over 75,000 classified military war logs by the USA relating to the war in Afghanistan; it is unaltered, unbiased and even in coverage. Its disclosure is unprecedented in the history of modern warfare and for the first time a detailed insider’s description of the day to day working of the world’s largest military power is openly available to public and academic scrutiny.

### 6.1.3 The Afghan War Diary

The c. 77,000 logs in the AWD are heavily detailed with each log indexed with its own spatial location as well as the day and time of occurrence. Additional fields describe the nature of the logged event such as *enemy action*, *friendly action* (subject to who initiated the action) or *explosive hazard*, give details of the events, and list the number of associated friendly, enemy and civilian casualties. To date, the data has been studied using

- data visualisation: chiefly carried out by journalists and researchers with a keen interest in the conflict. These approaches consist of simple visual inspections of the data such as plots of the number of logged events per month in different regions of Afghanistan for each event type (see the *github* page of Conway [233] for examples of such plots). Dewar, for instance, used non-parametric methods [234] to show the underlying report intensity in an animation [235]. The use of the same non-parametric methods for estimating the intensity can be seen also in [225]. Reporters typically concentrate on violent activity such as improvised explosive devices (IED) attacks which resulted in casualties and usually focus on showing the spatial distribution of such events [e.g. 236]. This high-level analysis of the data is good for providing quick overviews to the casual reader.

- descriptive statistics: carried out by academics with a professional interest in the subject. Approaches employed to date include i) fitting an anisotropic Gaussian distribution to the spatial distribution of logged events. Shifts in the mean and covariance of the resulting distribution across consecutive years is then used to effect conclusions on the evolution of the focal centre and spread of the zone of conflict. ii) space-time scan statistic clustering to extract space-time *hot spots* of logged events in each year. The locations of the clusters together with their change in distribution and sizes across consecutive areas are used to outline existing trends which are not immediately apparent from simple data inspection. Both methods i) and ii) are employed in [225, 231].

Data visualisation and descriptive statistics lack a means of analysing the spatiotemporal dynamic behaviour of the logged events in a rigorous manner. Additionally, without providing a *model* elucidating the development of the conflict state of affairs, they are unable to provide optimal statistical spatiotemporal inference of past behaviour (Was a highlighted spatiotemporal cluster a one off or was it representative of some underlying persisting *'conflict intensity'*?) and, more importantly, are unable to provide statistically founded predictions.

These limitations are obviated in the study proposed below. In what follows, preliminary analysis is first used to find a plausible descriptive model for the data in the Wikileaks data set (recall that an underlying infinite dimensional model is required in this framework). The modelling and inference mechanism of Chapter 4 is then used to carry out learning and prediction of the aforementioned model.

## 6.2 Preliminary analysis

Of the roughly 77,000 activity logs constituting the AWD, some are located outside Afghanistan's borders. These external events were omitted from the analysis bringing the actual number of logs considered down to 75,676. The following sections analyse the temporal and spatial content of this subset of the AWD.

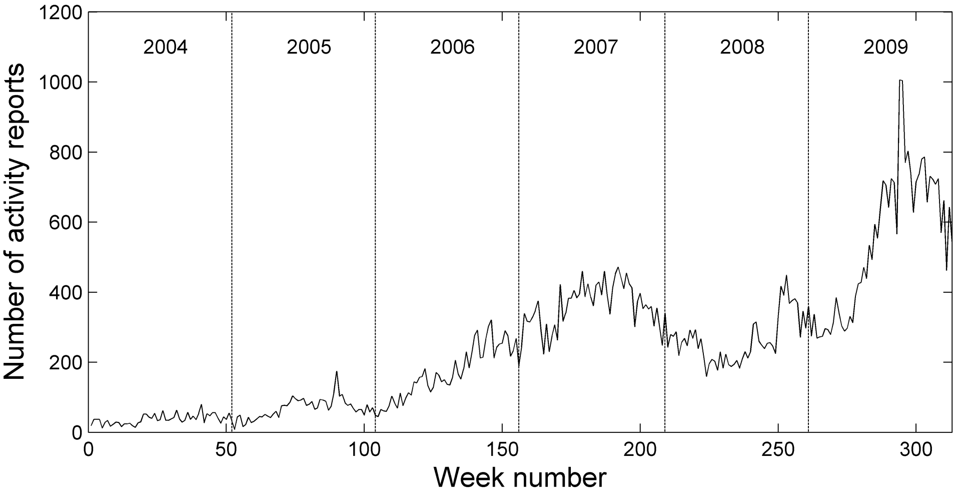


Figure 6.1: Weekly number of military activity reports in Afghanistan between January 2004 and December 2009 (bin size = 1 week).

### 6.2.1 Temporal analysis

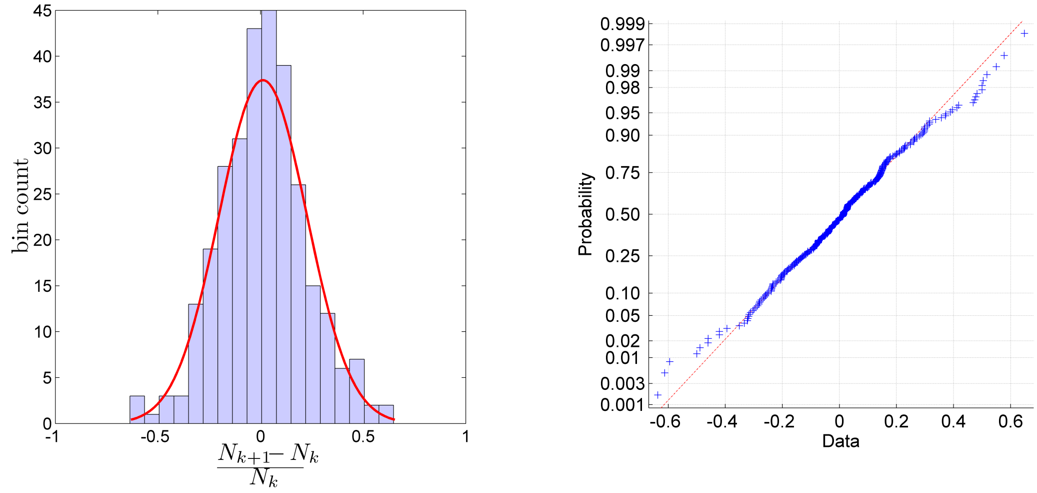
The first analysis stage of the AWD considers the temporal evolution, or sample path, of the overall report count per week as shown in Figure 6.1. This graph immediately shows strong evidence for an increase in activity over time and a strong peak in 2009 corresponding to the Afghan presidential election campaign (which, we recall, is not representative of what was reported in the media).

The behaviour of the data suggests a stochastic model with a positive exponential trend attributed to the steady increase of USA troops countered by the Afghan insurgency which can 'sustain itself indefinitely' [229]. Denote the number of logged events at week $k$ as $N_k$. As seen from Figure 6.2a and Figure 6.2b the quantity $(N_{k+1}-N_k)/N_k$ appears to be normally distributed. A two-tailed Shapiro-Wilks test for normality failed to refute the null hypothesis that these quantities were indeed generated by a normal distribution at the 10% level thus showing strong support for this observation. One may hence represent

$$\frac{N_{k+1}-N_k}{N_k} \sim \mathcal{N}(\tilde{r}\Delta_t, \sigma_w^2), \tag{6.1}$$

where $\Delta_t = 1$ week.

Another necessary test is that for *homoskedasticity*, which tests whether (6.1) experiences constant $\sigma_w^2$ across time. For this, the AWD was split into yearly intervals between 2004 and 2009. A Levene's test for heteroskedasticity was then used to check the null hypothesis that the variance across the different years was the same. The test



(a) (b)

Figure 6.2: (a) Histogram and fitted normal distribution to the fractional increments in log count per week in the AWD. (b) Normal probability plot of the fractional increments. The adherence of the data points (blue) to the straight line (red) is a strong indication of normality.

failed to reject it for the years 2006 to 2009 at the 10% level but not when including 2004 and 2005. The reason for rejection when including the earlier two years can be safely attributed to relatively low report count arising in noisy quantities when computing the fractional increments.

Preliminary analysis hence shows the data from the AWD exhibiting independent fractional changes with constant variance. Making the assumption that the report count is representative of an underlying spatiotemporal logging intensity $\lambda(t)$ which is generating the data, an intensity dynamic model can hence be given as

$$\mathrm{d}\lambda(t) = \tilde{r}\lambda(t)\mathrm{d}t + \sigma_w \lambda(t)\mathrm{d}\beta(t), \tag{6.2}$$

which is a *geometric Brownian motion*, a model widely applied in finance and in many other application areas [237]. The quantity $\tilde{r}$ is commonly known as the percentage drift and $\sigma_w$ as the volatility. It is required to put the intensity dynamic equation (6.2) into the exponential form

$$\lambda(t) = \exp(\mu + x(t)), \qquad \ln \lambda(t) = \mu + x(t), \tag{6.3}$$

for application of the theory developed in Chapters 3 and 4. This entails the use of the following lemma.

**Lemma 6.1 (Ito's lemma [104, Section 4.5])** *Let $\lambda(t)$ be the unique solution of the SDE*

$$d\lambda(t) = f(\lambda(t), t)\,dt + g(\lambda(t), t)\,d\beta(t). \tag{6.4}$$

*Let $\psi(\lambda(t), t)$ be i) continuously differentiable in $t$ and ii) have continuous second order partial derivatives with respect to $\lambda(t)$. Then $\psi$ is governed by the SDE*

$$d\psi = \frac{\partial \psi}{\partial t} dt + \frac{\partial \psi}{\partial \lambda} d\lambda + \frac{1}{2} g(\lambda(t), t)^2 \frac{\partial \psi^2}{\partial \lambda^2} dt. \tag{6.5}$$

Applying Lemma 6.1 to (6.2) with $\psi(\lambda(t), t) = \ln \lambda(t)$, $f(\lambda(t), t) = \tilde{r}\lambda(t)$ and $g(\lambda(t), t) = \sigma_w \lambda(t)$ one obtains

$$\begin{aligned} \mathrm{d}(\ln \lambda(t)) &= \frac{1}{\lambda}\mathrm{d}\lambda - \frac{1}{2}\sigma_w^2 \mathrm{d}t \\ &= \left(\tilde{r} - \frac{1}{2}\sigma_w^2\right) \mathrm{d}t + \sigma_w \mathrm{d}\beta(t). \end{aligned} \tag{6.6}$$

Without loss of generality, letting $r = \tilde{r} - \sigma_w^2/2$,

$$\mathrm{d}(\ln \lambda(t)) = r\mathrm{d}t + \sigma_w \mathrm{d}\beta(t). \tag{6.7}$$

Comparing (6.7) to (6.3), since $\mu$ is a constant ($\mathrm{d}\mu = 0$), the underlying evolution equation of $x(t)$ is hence given by

$$\mathrm{d}x(t) = r\mathrm{d}t + \sigma_w \mathrm{d}\beta(t). \tag{6.8}$$

The arguments presented above readily extend to the general spatiotemporal case. Taking $z(t) \in H$ as the hidden spatiotemporal state and $\lambda(t) \in H$ as the spatiotemporal intensity one obtains the system of equations

$$\mathrm{d}z(t) = r\mathrm{d}t + \sigma_w \mathrm{d}W(t), \tag{6.9}$$

$$\lambda(t) = \exp(\mu + z(t)), \tag{6.10}$$

from where it is implicitly assumed that

$$\mathrm{d}\lambda(t) = \left[r + \frac{1}{2}\sigma_w^2\right] \lambda(t)\mathrm{d}t + \sigma_w \lambda(t)\mathrm{d}W(t). \tag{6.11}$$

Here $r \in H$ is a heterogeneous temporally independent spatial growth rate, $\mu$ is a background event rate (with the same role as in the neural firing model) and $W(t)$ is an

$H$-valued Wiener process with smooth covariance operator $Q : H \to H$. The intensity function $\lambda(t)$ is then said to follow a *spatiotemporal geometric Brownian motion* with percentage drift $r + \sigma_w^2/2$ and spatially constant percentage volatility $\sigma_w$. The model allows for different provinces in Afghanistan to experience different growth rates (for instance the growth in Helmand and Kandahar was overall much more than in other provinces such as Herat) but assumes that the volatility or *risk* is constant nationwide. For simplicity the background event rate is also assumed to be constant nationwide.

### 6.2.2 Spatial analysis and basis function placement

An implicit assumption made in (6.9) is that the spatial interactions are negligible and that the system is spatially uncoupled. This assumption needs to be verified using tools emerging from point process statistics.

The key components required for this analysis are the PACF of Definition 4.1

$$g_{k,k}(\boldsymbol{s},\boldsymbol{r}) = \frac{\lambda_{k,k}^{(2)}(\boldsymbol{s},\boldsymbol{r})}{\lambda_k^{(1)}(\boldsymbol{s})\lambda_k^{(1)}(\boldsymbol{r})}, \tag{6.12}$$

and the *pair cross-correlation function (PCCF)*

$$g_{k,k+1}(\boldsymbol{s},\boldsymbol{r}) = \frac{\lambda_{k,k+1}^{(2)}(\boldsymbol{s},\boldsymbol{r})}{\lambda_k^{(1)}(\boldsymbol{s})\lambda_{k+1}^{(1)}(\boldsymbol{r})}, \tag{6.13}$$

where recall that $\lambda_k^{(1)}(\boldsymbol{s}) = \mathbb{E}[\lambda_k(\boldsymbol{s})]$, $\lambda_{k,k}^{(2)}(\boldsymbol{s},\boldsymbol{r}) = \mathbb{E}[\lambda_k(\boldsymbol{s})\lambda_k(\boldsymbol{r})]$ and where $\lambda_{k,k+1}^{(2)}(\boldsymbol{s},\boldsymbol{r}) = \mathbb{E}[\lambda_k(\boldsymbol{s})\lambda_{k+1}(\boldsymbol{r})]$. Now, let $\Delta_{W_k(s)} \sim GP(0, k_Q)$ be a Gaussian process with mean function 0 and covariance function $k_Q$; then the emerging form of $z(t)$ in (6.9), and in particular its discretised version (using an explicit Euler scheme)

$$z_{k+1}(\boldsymbol{s}) = z_k(\boldsymbol{s}) + r\Delta_t + \Delta_{W_k(s)}\Delta_t, \tag{6.14}$$

implies that at each week $k$, $z_k(\boldsymbol{s})$ is a realisation from a GP, say $z_k(\boldsymbol{s}) \sim \mathcal{GP}(\hat{z}_k(\boldsymbol{s}), \sigma_k^2\psi_k(\boldsymbol{s},\boldsymbol{r}))$, where $\psi_k$ is a correlation function with $\psi_k(\boldsymbol{s},\boldsymbol{s}) = 1$. Therefore it follows that $\lambda_k(\boldsymbol{s})$ as defined in (6.10) in this case is a LGCP of Definition 4.2. Recall that from Lemma 4.1

$$g_{k,k}(\boldsymbol{s},\boldsymbol{r}) = \exp(\sigma_k^2\psi_k(\boldsymbol{s},\boldsymbol{r})), \tag{6.15}$$

leading to the following theorem:

**Theorem 6.1** *For evolving GPs following the spatially uncoupled dynamics of (6.14)*

$$g_{k,k}(\boldsymbol{s},\boldsymbol{r}) = g_{k,k+1}(\boldsymbol{s},\boldsymbol{r}). \tag{6.16}$$

*Proof.* Consider the quantity

$$\begin{aligned}\mathbb{E}[\lambda_k(\boldsymbol{s})]\mathbb{E}[\lambda_{k+1}(\boldsymbol{r})] &= \mathbb{E}[\exp(\mu + z_k(\boldsymbol{s}))]\mathbb{E}[\exp(\mu + z_k(\boldsymbol{r}) + r\Delta_t + \Delta_{W_k(s)}\Delta_t)] \\ &= \exp(2\mu + r\Delta_t)\mathbb{E}[\exp(z_k(\boldsymbol{s}))]\mathbb{E}[\exp(z_k(\boldsymbol{r}))]\mathbb{E}[\exp(\Delta_{W_k(s)}\Delta_t)],\end{aligned} \tag{6.17}$$

where the expectations could be factorised as a result of the additive disturbance at $k$ being independent of $z_k$. The second moment is given by

$$\begin{aligned}\mathbb{E}[\lambda_k(\boldsymbol{s})\lambda_{k+1}(\boldsymbol{r})] &= \mathbb{E}[\exp(\mu + z_k(\boldsymbol{s}) + \mu + z_k(\boldsymbol{r}) + r\Delta_t + \Delta_{W_k(s)}\Delta_t)] \\ &= \exp(2\mu + r\Delta_t)\mathbb{E}[\exp(z_k(\boldsymbol{s}) + z_k(\boldsymbol{r}))]\mathbb{E}[\exp(\Delta_{W_k(s)}\Delta_t)],\end{aligned} \tag{6.18}$$

so that

$$\begin{aligned}g_{k,k+1}(\boldsymbol{s},\boldsymbol{r}) &= \frac{\mathbb{E}[\exp(z_k(\boldsymbol{s}) + z_k(\boldsymbol{r}))]}{\mathbb{E}[\exp(z_k(\boldsymbol{s}))]\mathbb{E}[\exp(z_k(\boldsymbol{r}))]} \\ &= \frac{\exp(\hat{z}_k(\boldsymbol{s}) + \hat{z}_k(\boldsymbol{r}) + \sigma_k^2 + \sigma_k^2\psi(\boldsymbol{s},\boldsymbol{r}))}{\exp(\hat{z}_k(\boldsymbol{s}) + \hat{z}_k(\boldsymbol{r}) + \sigma_k^2)} \\ &= \exp(\sigma_k^2\psi(\boldsymbol{s},\boldsymbol{r})) = g_{k,k}(\boldsymbol{s},\boldsymbol{r}),\end{aligned} \tag{6.19}$$

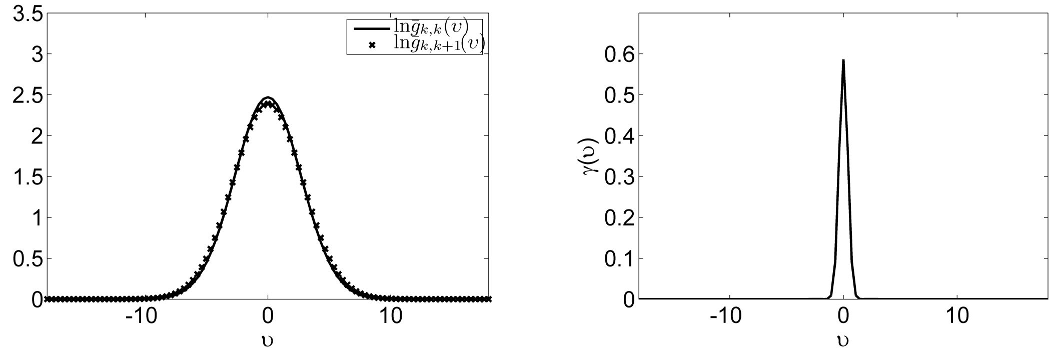
where the last equality follows from Lemma 4.1. ■

The key insight of Theorem 6.1 is that the relationship between the log PCCF and the log PACF can be used to detect spatial dynamic interactions across different time steps. The approach presented is also directly applicable to SIDE type models where the relationship would also be indicative of the transition kernel shape. Since this is beyond the scope of the present argument it shall not be discussed further.

If the hypothesis of spatially uncoupled dynamics is correct, then the PACF and the PCCF as obtained from the data should be equal, or, letting $\upsilon = ||\boldsymbol{s} - \boldsymbol{r}||$

$$\gamma(\upsilon) = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\ln g_{k,k+1}(\upsilon))}{\mathcal{F}(\ln g_{k,k}(\upsilon))}\right) = \delta(\upsilon). \tag{6.20}$$

For nonparametric estimation of the quantities $g_{k,k+1}(\upsilon)$ and $g_{k,k}(\upsilon)$ the reader is referred once again to Appendix C. The results of carrying out correlation tests on the raw data



(a)

(b)

Figure 6.3: (a) Average log PACF $\ln \bar{g}_{k,k}(\upsilon)$ and average log PCCF $\ln \bar{g}_{k,k+1}(\upsilon)$. (b) $\gamma(\upsilon)$ as computed from (6.20).

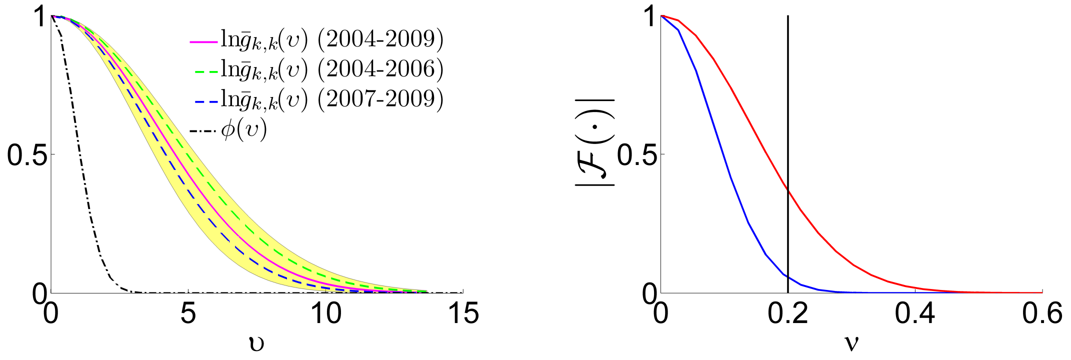
were very revealing and are summarised in Figure 6.3.[2] Figure 6.3a shows that on average the log PACF is equal to the log PCCF and that $\gamma(\upsilon)$, as computed from (6.20), is very narrow in relation to the extent of the spatial correlations in the field (recall that the log PACF is proportional to the correlation function). This implies that $\gamma(\upsilon)$ may be safely approximated to a delta function, corresponding to negligible spatial interactions across adjacent time frames.

The temporal analysis together with the spatial analysis thus shows strong support for the set of equations (6.9) and (6.10) describing a reasonable model for the dynamic behaviour of the AWD. Note that (6.9) is not an SPDE since $r$ is not a differential operator. Nonetheless the approach considered in this thesis readily extends to this case, and this example shows that the developed methodology may be applied to a broader class of infinite-dimensional stochastic equations, and not only to linear equations incorporating differential operators.

The next step before carrying out joint field-parameter inference is the decomposition of the growth function $r$ and the underlying spatiotemporal field using basis functions.

**Basis function selection:** The frequency content of the field is found from the autocorrelation function by Theorem 4.2 which in turn is determined from the PACF through Lemma 4.1. Interestingly, as shown in Figure 6.4a, the PACF over the 6 year period of the AWD does not change significantly, an indication of clusters which do not change

[2] For convenience the original spatial domain of the AWD denoted in latitude/longitude coordinates was mapped on to a symmetric space with the bottom left-hand corner on the origin using a linear map $\mathcal{L}u = [\rho_1(u_1 - \tau_1), \rho_2(u_2 - \tau_2)]$, where $u_1$ and $u_2$ denote the longitude and latitude coordinates in real space. The scale/shift parameters were chosen to be $\tau = [57.5, 28.5]$ and $\rho = [2, 3.3]$.



(a)

(b)

Figure 6.4: (a) Average log PACF with respect to relative distance $\upsilon = ||\boldsymbol{s} - \boldsymbol{r}||$ for different time intervals (blue, magenta, green), the 95% spread of the log PACF across all the years (yellow band) and the selected isotropic basis function (black) which is clearly narrow enough (of sufficient high frequency) to represent the spatial correlations in the field. (b) Spatial Fourier transform of underlying field (blue) and of $\phi(\upsilon)$ (red). The vertical black line denotes the selected cutoff frequency of $\nu_c$ =0.2 cycles/unit.

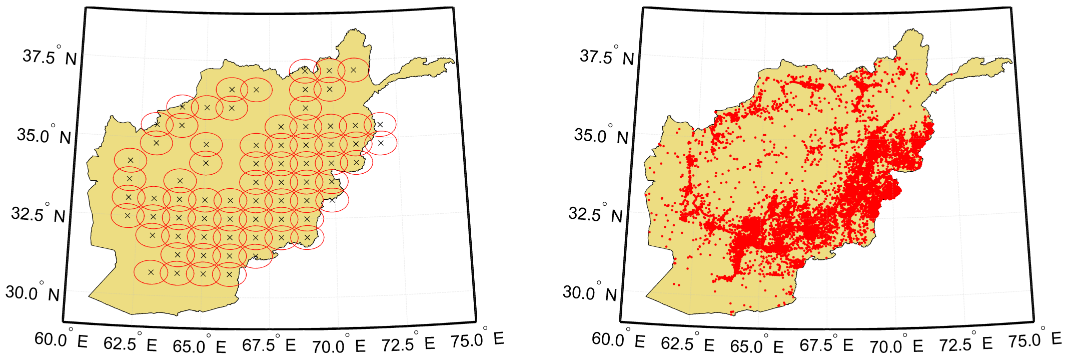
in size in time. This somewhat corroborates a graphical inspection of Figure 12 in [225] depicting clusters obtained using the Getis-Ord clustering statistic.

From the frequency response in Figure 6.4b a cutoff frequency of $\nu_c = 0.2$ cycles/unit was selected (see Footnote 2 on pg. 175) for computation of the basis functions width. From (3.26) this is given by

$$\sigma_b^2 = \frac{1}{2\nu_c^2\pi^2} = 1.27, \tag{6.21}$$

rounded down to $\sigma_b^2 = 1$ so that the corresponding local GRBF $\tau$ is given by $\tau = \sqrt{\pi}$. The isotropic basis function used in the analysis is depicted in Figure 6.4a.

Initially basis functions were placed on a 16×16 grid with an inter-centre spacing of $\Delta_s = 1.9$ corresponding to an oversampling parameter in (3.21) of $\alpha_0 = 1.32$ and covering the whole of Afghanistan. Many of these basis functions were however considered redundant, representing areas exhibiting no logged events, or a very small number of events. To avoid problems of identifiability in these regions (see Section 4.1.3, pg. 98), the constant background intensity baseline $\mu$ was used to represent activity in these areas. Each basis function was analysed separately; if a basis function had its centre more than 0.4 spatial units outside Afghanistan or had on average less than 8 logged events per year within 1.3 standard deviations from its centre (corresponding to $\mu = -3.5$) it was omitted. The final arrangement of the basis functions together with the spatial distribution of all the logged events of the AWD in Afghanistan is shown in Figure 6.5a



(a) (b)

Figure 6.5: (a) Basis functions in spatial domain where the ‘x’ denotes the function centre and the circle the 1.3 × standard deviation interval. (b) Spatial distribution of activity logs in the AWD between 2004 and 2009.

and Figure 6.5b respectively. Note that basis functions are omitted in ‘quiet’ areas. As can be seen from (6.10), the intensity where there are no basis functions simply reduces to $\exp(\mu)$.

## 6.3 Model decomposition and VB inference

Applying the Galerkin reduction method of Theorem 3.1 to (6.9) one obtains the finite dimensional representation

$$\mathrm{d}\mathbf{x}(t) =\tilde{\mathbf{r}}\mathrm{d}t + \sigma_w \mathbf{\Psi}_x^{-1}\mathrm{d}\boldsymbol{\beta}(t), \tag{6.22}$$

where $\tilde{\mathbf{r}} = \mathbf{\Psi}_x^{-1}\boldsymbol{\psi}_r$, $\boldsymbol{\psi}_r = \langle r, \boldsymbol{\phi}\rangle$, $\boldsymbol{\phi} = [\phi_1, \phi_2, \dots, \phi_n]^T$ and $\boldsymbol{\beta}(t)$ is a vector of correlated time series’ whose increments are normally distributed with 0 mean and covariance matrix $\boldsymbol{Q}$. Subsequent application of the Euler scheme gives

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta_t\tilde{\mathbf{r}} + \mathbf{w}_k, \tag{6.23}$$

where $\mathbf{x}_k := \mathbf{x}(k\Delta_t)$ and where $\mathbf{w}_k \in \mathbb{R}^n$ is additive temporally white Gaussian noise with zero mean and covariance $\mathbf{\Sigma}_w = \sigma_w^2\Delta_t\mathbf{\Psi}_x^{-1}\boldsymbol{Q}_n\mathbf{\Psi}_x^{-1}$. To cater for heterogeneity in the form of a spatially varying growth rate, it is required to also project the growth function $r$ onto $H_n$ so that $r = \sum_{i=1}^n \phi_i\theta_i$; it then follows that $\tilde{\mathbf{r}} = \boldsymbol{\theta}$. This gives a fully decoupled model

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta_t \boldsymbol{\theta} + \mathbf{w}_k, \tag{6.24}$$

$$\lambda_k = \exp(\mu + \boldsymbol{\phi}^T \boldsymbol{x}_k). \tag{6.25}$$

The unknown quantities in this model are the underlying states $\mathcal{X} = \{\boldsymbol{x}_k\}_{k=0}^K$ composing the spatiotemporal field which gives optimal smoothed logging intensity quantities, and the unknown parameters $\boldsymbol{\theta}$ which compose the growth function. The model (6.24) and (6.25) is slightly different from that considered in Chapter 4 however the construction of a smoother proceeds on the same lines. In particular the required algorithm for the forward message is identical to that of SPDEs as shown in Algorithm A.3 with $\boldsymbol{A} = \boldsymbol{I}$ and with the following changes

- In the forward pass the quantity $\tilde{\boldsymbol{x}}_k$ is computed as

$$\begin{aligned}\tilde{\boldsymbol{x}}_k = \tilde{\boldsymbol{\Sigma}}_k \Big[ \sigma_w^{-2} \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Sigma}_{k-1}^* (\boldsymbol{\Sigma}_{k-1|k-1}^{-1} \hat{\boldsymbol{x}}_{k-1|k-1} \\ - \sigma_w^{-2} \widetilde{\boldsymbol{Q}}^{-1} \mathbb{E}[\boldsymbol{\theta}] \Delta_t) + \sigma_w^{-2} \widetilde{\boldsymbol{Q}}^{-1} \mathbb{E}[\boldsymbol{\theta}] \Delta_t \Big].\end{aligned} \tag{6.26}$$

- As a result of removing basis functions in regions of low observed logs, the ill-conditioning of $\tilde{\beta}(\boldsymbol{x}_k)$ as discussed in Section 4.1.3 is no longer an issue and (un-modified) Laplace approximations of the backward message may be computed. The algorithm for the backward message is given in Algorithm A.5.

Note that computation of the smoothed estimates remains unchanged and the evaluation of the cross-covariance is not required for this model. The variational parameter posterior at the $(i+1)^{th}$ iteration $\tilde{p}(\boldsymbol{\theta})^{(i+1)}$ is given as

$$\begin{aligned}\tilde{p}(\boldsymbol{\theta})^{(i+1)} \propto p(\boldsymbol{\theta}) \exp \Bigg( -\frac{1}{2} \mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}} \Bigg[ \sum_{k=0}^{K-1} (\boldsymbol{x}_{k+1} - \boldsymbol{x}_k - \boldsymbol{\theta}\Delta_t)^T \sigma_w^{-2} \widetilde{\boldsymbol{Q}}^{-1} \\ \times (\boldsymbol{x}_{k+1} - \boldsymbol{x}_k - \boldsymbol{\theta}\Delta_t) \Bigg] \Bigg),\end{aligned} \tag{6.27}$$

to give $\boldsymbol{\theta}^{(i+1)} \sim \mathcal{N}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ where

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \left[ \boldsymbol{\Sigma}_{\boldsymbol{\theta},p} \hat{\boldsymbol{\theta}}_p + \Delta_t \sigma_w^{-2} \widetilde{\boldsymbol{Q}}^{-1} \sum_{k=0}^{K-1} (\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}} [\boldsymbol{x}_{k+1} - \boldsymbol{x}_k]) \right], \tag{6.28}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = (\boldsymbol{\Sigma}_{\boldsymbol{\theta},p}^{-1} + \Delta_t^2 K \sigma_w^{-2} \widetilde{\boldsymbol{Q}}^{-1})^{-1}. \tag{6.29}$$

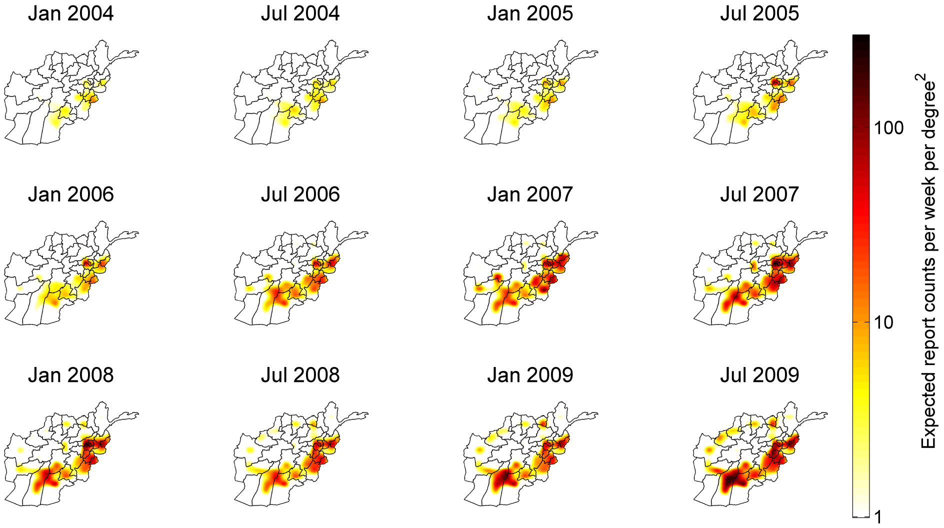


Figure 6.6: Estimated mean intensity $\mathbb{E}[\lambda_k(\boldsymbol{s})]$ on the first week of the month and respective year.

The algorithm was run with $\mu = -3.5$ corresponding to the activity baseline used for basis function retention, and $\sigma_w = 0.2$. The latter was chosen to be equal to the standard deviation of the increments (6.1) in 2006, the largest of the four years 2006-2009 for which homoskedasticity was met. The approximate inference methodology allowed a considerably fast and efficient joint estimation of the intensity of logged events and underlying parameters composing the growth function; despite the large number of unknown states and parameters, inference was complete in only about six hours on a standard personal computer.[3] This performance is very encouraging when considering the size of the data set under investigation.

## 6.4 Results and discussion

Figure 6.6 shows the temporal progression of the mean posterior intensity function underlying the AWD events at all places in Afghanistan on the first week of a given month and year. Note how the intensity is naturally smoothed by the algorithm, which optimally interpolates future and past observations (within the training period for which data is available); this contrasts sharply with visualisations which ignore the dynamics and which would yield abrupt changes in intensity if applied at the fine time-scale of 1

[3] With Intel®Core$^{\text{TM}}$2Duo E8400 @ 3.00GHz, 2GB RAM.

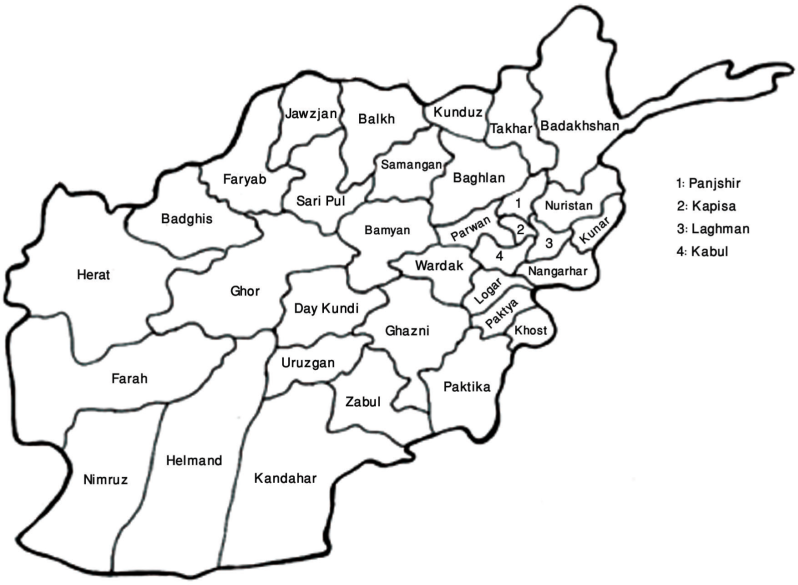


Figure 6.7: Provincial map of Afghanistan accurate as of 2010.

week.

The figure provides an intuitive visualisation which captures important geographical features of the war scenario. Regions of high intensity in the most recent months include Sangin in northern Helmand (see Figure 6.7), one of the most dangerous places in Afghanistan, notorious for the thousands of IEDs and frequent suicide bombings [238]. Other regions, such as Kabul, Nangarhar and Paktya (the latter two close to the Pakistan border) on the other hand have witnessed relatively high activity all throughout the six-year interval. Also very apparent is the emergence of a high intensity ring starting from Kabul extending downwards towards Kandahar, up through Herat, through Mazar-E Sharif in the Balkh province and back to Kabul. This roughly elliptical shape corresponds to the country's 'ring road', commonly targeted by insurgent activity and IED placement [225].

A major advantage of the model-based approach is the ability to establish quantitative conclusions on activity growth rates between 2004 and 2009 at a glance from the mean growth rate estimate (Figure 6.8A). One may distinguish between *event hot spots* (i.e. regions with clustered events), and *growth hot spots* (events with rapid growth in activity). While some of the high growth areas, such as Helmand, also had an overall high count of events, this is not the general case; for example, Sar-e Pul and Balkh in the North and the Badghis province in the West both have witnessed a modest number of total events but have shown significant growth in activity in recent years. Note that such patterns are a result of viewing the conflict dynamics from a systems perspective; no further analysis is needed to obtain this spatiotemporal behaviour on a provincial level (Figure 6.8B-F). In the insets it is seen how solely conditioned on the a posteriori estimated intensity in January 2004, the model predicted output (red) in most cases accurately models the observed trend. One exception is Kabul which sees a sharp rise in activity in 2007. The smoothed intensity (green) provides a good tracking of the observed log count in every case.

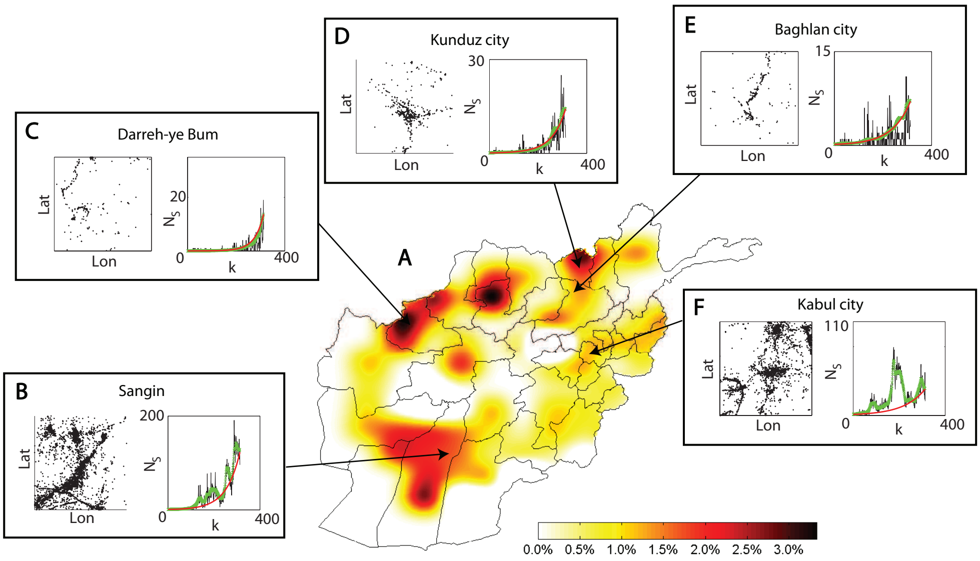


Figure 6.8: (**A**) Posterior mean fractional increase in activity logs per week in Afghanistan between 2004 and 2009. Only regions with positive overall growth are shown. (**B - F** left) Spatial map of all events occurring in a square of side 100km centred on the city under study. (**B - F** right) Number of weekly events $N_S$ at week $k$ in these regions (black) together with the a posteriori estimated count (green) and the model predicted count conditioned only the estimated intensity in January 2004 (red).

**Prediction:** AOG initiated attacks (as reported by ANSO) in 2009 continued to increase from 2008, with 8 provinces reporting an escalation of more than 100% (corresponding to a growth factor of 2). A fundamental question to ask is whether the reported figures could have been of any use in predicting the provincial growth in AOG initiated attacks in 2010 as a forewarning to the local population and stationed military. However a simple correlation analysis indicated that the growth in 2009 was not a good indication of what happened in 2010; even with 4 clear province outliers omitted from the analysis

(Sari-Pul, Bamyan, Paktya and Ghazni), the correlation coefficient was found to be only 0.34 at a two-tailed $p$-value of 0.08 (not significant at the 5% level). Further, the simple use of the activity growth in 2009 as a predictor for 2010 (denoted as *ANSO prediction*) is unable to provide credibility intervals which would undoubtedly be of considerable use in decision making and strategy planning. Could the growth model reproduced above be used to predict the AOG initiated activity in 2010 (denoted here as *model prediction*)? This proposal may not be too farfetched, given that there exists a positive correlation between violent activity and the number of logged events in the AWD [225] and the inferred model is statistically optimal in some sense. It is important to bear in mind that the AWD only contains data until 2009; any data in 2010 may be treated as validation data.

The procedure adopted was as follows:

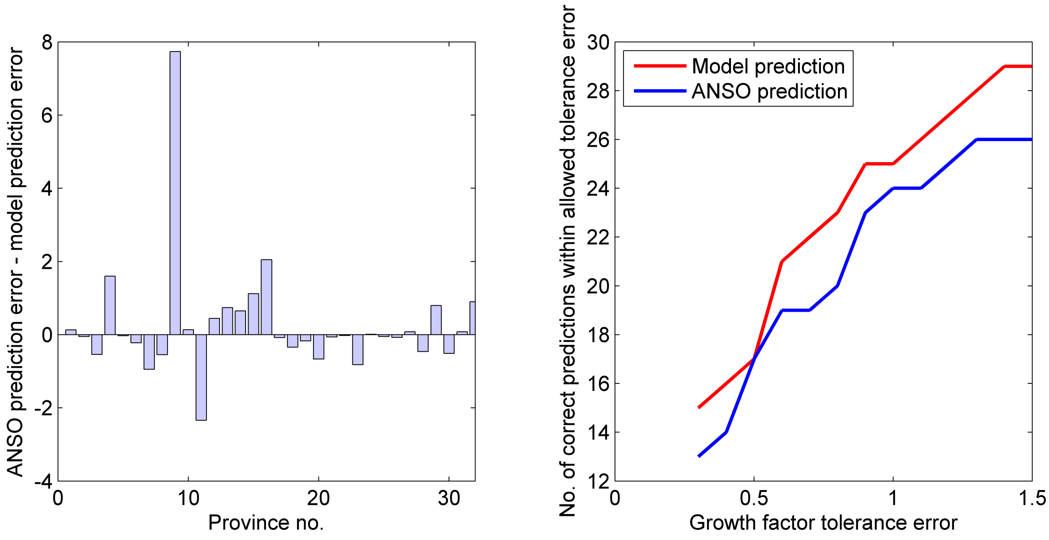
1. A linear relationship was established between the 2008/2009 provincial growth of AOG initiated attacks as reported by ANSO and the corresponding provincial growth in the Wikileaks data number of logged events.

2. 2,000 MC forward simulations runs were carried out using the growth model (6.9) and (6.10) in order to obtain the mean and credibility intervals for the number of logged events in 2010.

3. The linear relationship established in (1) was used to map the predicted growth in the Wikileaks number of logged of events to the predicted growth of AOG initiated attacks in 2010.

The results are very encouraging. Even when treating the mean value of the model predicted growth in AOG attacks as the predicted growth, the generative model is seen to produce better estimates than what analysts would typically obtain just from events in the previous year (growth in 2009). The difference between the ANSO prediction error and the model prediction error is seen in Figure 6.9a where it is clear that the growth model on average performs better than the naïve prediction method. This observation is further confirmed in Figure 6.9b where it is seen that for a given tolerance of error, the proposed approach consistently predicts correctly the growth activity in more provinces than the ANSO prediction does. The model predictor allows for up to 20% less tolerance error for the same number of provinces with correctly predicted growth factor.

The model predictions exhibited relatively large errors in only 4 out of the 32 provinces; growth in Badghis, which had been alarming in recent years [239, 240], was dramatically overestimated, while the growth in 3 neighbouring provinces close to the



(a) (b)

Figure 6.9: (a) Difference in ANSO prediction with model prediction when considering the mean value from the forward MC runs. Positive quantities indicate provinces where the model predictor performed better than the ANSO predictor. (b) Cumulative graph of correct predictions within a growth tolerance error.

Pakistani border (Ghazni, Pakitka and Paktya) was underestimated. This unusual pattern in these three neighbouring provinces was highlighted in the ANSO report, which described it as a highly unusual 'surprise' deterioration [228]. After the removal of the four outliers, a correlation analysis indicated a correlation coefficient of 0.72 between the mean prediction and the reported growth, at a $p$-value of 0.0001. This compares very favourably with the correlation coefficient of 0.34 obtained when comparing the 2009 growth with that in 2010 (also after removal of four outliers).

The mean of the predicted growth, however, exhibits large errors so that its use is limited in decision making. The most important feature of the proposed approach for conflict analysis is in fact a provision of associated credibility intervals with the predictions. As seen in Figure 6.10 in 75% of the provinces (24 out of 32) the recorded growth by the ANSO lay within the three-sigma confidence intervals of the predictions (as generated by MC runs). The model managed to, for instance, predict the high increase in growth in Sar-e-Pul but also outlined that one cannot really be sure whether the growth will be anywhere between 100% and 500% (it turned out to be 250%). On the other hand for most provinces the confidence intervals were more useful; it said that the growth in Helmand will be somewhere in between 100% and 170% and the true growth was 124%.

Of the eight incorrect estimates, six were underestimates and two were overestimates.

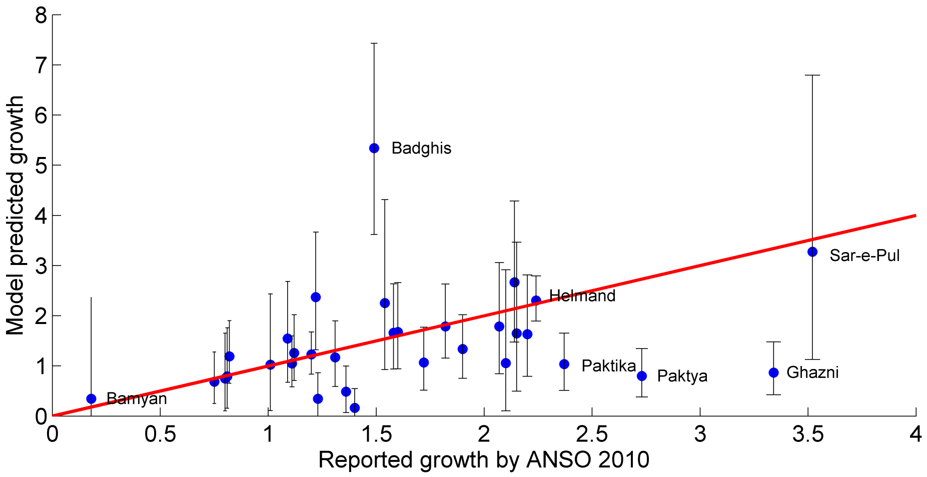


Figure 6.10: Comparing the reported growth by ANSO with that predicted from the generative model using Wikileaks data where the blue marks denote the mean predicted growth and the bars correspond to three-sigma confidence intervals. The red line corresponds to exact prediction. 75% of the predictions give credibility intervals which enclose the true reported growth rate.

Interestingly all of the six provinces neighbour each other close to the Pakistani border (see Figure 6.11). Such a spatial correlation in the error is indicative of an external influence in this region of Afghanistan causing a violation of the model, such as increased military presence or significant opposition (as supported by ANSO). Four of the eight provinces were considered outliers when studying the prediction means (Badghis, Ghazni, Paktika and Paktya).

Predictions of this quality are surprising in a simple model applied to an enormously complex system (with heavy external influences). It is important to remark that these predictions are accurate based solely on the historical sequence of events. If these type of predictions can be made using solely one data set, without all the other inside knowledge available to security organisations, one may only presume that much more accurate predictions with more resources at hand are indeed possible. It is thus highly likely that the proposed method, combined with other knowledge, such as election dates and pre-planned strategic military activity, will yield predictions of an even higher quality than those demonstrated here. Another advantage of the proposed method is of course that incorporation of further prior knowledge is somewhat straightforward through the use of priors in the Bayesian framework employed.

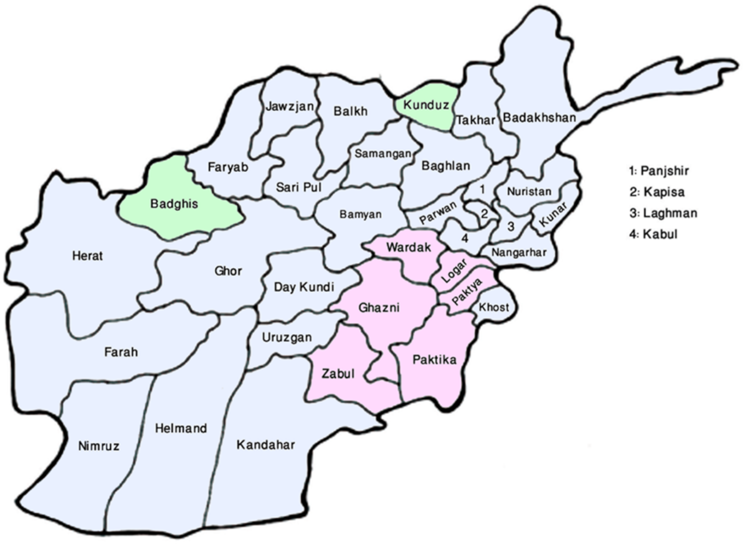


Figure 6.11: Provincial map in which AOG activity was within predicted three-sigma confidence intervals (blue), significantly more than expected (magenta) and significantly less than expected (green).

## 6.5 Conclusion

Whilst today many social scientists advocate the use of computational methods in their field [7], surprisingly, the use of such tools in conflict analysis is still in its early stages.

> "I would say the weather guys are far ahead of where we are" *V. Subrahmanian, co-director of the Laboratory for Computational Cultural Dynamics, 2011 [230].*

Nonetheless, the potential of sophisticated computational methods in this field is evidenced by the treatment in this chapter. The methodology adopted is inherited from the developments of Chapter 3 and 4, which, as expected from previous simulation studies was fast and efficient.

The intensity of the AWD was shown to follow a geometric Brownian motion with little or no spatial dynamic interactions. In essence the work presented here opposes the claim that 'the current generation of models is [not] telling people anything an expert in the relevant subject wouldn't already know' [230]; it has provided further evidence of an emerging simple behaviour previously gone unnoticed in a conflict scenario [8]. It is envisaged that this approach may be easily extended to more complicated conflict scenes; such as those containing diffusions (for instance across provincial boundaries) or drifts.

In this work the data in the AWD was treated as a univariate point process. However the data is excellently suited for multivariate point process models where the different *marks* would be corresponding data labels, such as *friendly action* or *explosive hazard*. It would be interesting to study the dynamics of the individual variates and the dynamic interactions between the variates. Another interesting study would consider the IED incidence along the country's ring-road (on a one-dimensional spatial domain) where spatial interactions are likely to be observed. A final point is that the data was considered on a nationwide scale; it is likely that high frequency dynamics are observed when studying a single province (for instance Kabul or Helmand, which contain many events). The provinces may then be modelled using the same approach as above; the process of superimposing the micro-scale model on the macro-scale would then result in a multi-resolution modelling approach.

In the interest of brevity, the online version of the algorithm was not implemented, however from Chapter 5 it is easy to see that this would not require substantial extra effort from an algorithmic perspective and for the batch DVBF a plausible value of $L$ would be $L = 1$ year. Other than this, this chapter demonstrates that the proposed approach satisfies all the objectives laid out in Section 1.3.2 since it i) infers a spatiotemporal field from point process observations, ii) extracts heterogeneous characteristics of the field, iii) maintains a continuous-space representation and iv) employs estimation methods which are approximate, but fast and efficient. The next chapter will put this contribution into perspective and discuss paths for future work.

# Chapter 7

# Conclusion

This thesis has presented the development of an integrated systems theoretic framework for the study of spatiotemporal systems. Pivotal to this framework is a methodology developed for linear systems following infinite-dimensional SDEs and in particular SPDEs; a continuous-space continuous-time class of models which easily describes spatially heterogeneous dynamics. The approach maintains a continuous-space representation throughout and employs fast and efficient inference mechanisms, offline or online, from both continuous and point process observations. Summarily, the adopted methodology consists of:

- placing the infinite dimensional SDE into a finite dimensional discrete-time state-space model, with the ensuing dimensionality decoupled from the dimensionality of the observation process,
- employing frequency analysis to obtain a set of local basis functions from both continuous observations and point process observations, and
- estimating both the field and the parameters of the spatiotemporal system from the set of observations.

The developed methods were shown to perform well using both synthetic data and real data sets and thus all objectives set out for in Section 1.3.2 have been achieved.

## 7.1 Summary

Chapter 2 discusses different spatiotemporal model classes which are in common use and justifies why the study of SPDEs is warranted in this work. Some theoretic properties pertaining to SPDEs are also highlighted; these properties, inherited from the mathematical literature on SDEs, provide a foundation for the model reduction approach employed in later chapters. The chapter proceeds to study prevalent inference mechanisms for state-space identification and shows clear potential advantages of VBEM over EM and MCMC methods. In particular, it is seen that VBEM is deterministic (fast), it retains uncertainty measures and approximates distributions around the mean of the true posterior rather than the mode. The latter property was envisioned to be highly useful when dealing with skewed distributions such as in the case of point process systems.

The framework described in Chapters 3 and 4 is a key novelty of this work. It describes an approach for taking a complex, large-scale problem and breaking it down into smaller, more manageable tasks, as described in the preface to this chapter. These chapters also provide solutions and methods for each of these tasks.

In Chapter 3 the EM, VBEM and Gibbs sampler were implemented for the estimation of SPDEs which can be expressed linearly in the parameters. All methods were compared to one another; it was noted that in real-data applications Gibbs sampling and other MCMC techniques are likely to be unsatisfactory due to computational issues. An advantage of VBEM over EM is that it provides confidence bounds on the parameters; however, as remarked by comparing to the robust parameter uncertainty estimator of Duan, these bounds may not be physically significant. The VBEM algorithm is seen to provide inadequate uncertainty measures through its use of the complete-data information matrix for variance estimation, a quantity which varies significantly from the observed-data information matrix in noisy conditions.

Chapter 4 presents an entirely new approach to dealing with spatiotemporal point processes which, traditionally, are studied using only geostatistical models. The framework is the same in principle as that of Chapter 3; it consists of finding a state-space representation of the underlying SPDE, finding a set of basis functions, and estimating the parameters. The latter two steps are more involved in the context of point processes. Basis function placement is carried out using nonparameteric estimation methods in order to obtain the average frequency content of the signal. Parameter estimation is carried out using a VB-Laplace scheme, an approach which still favours the *mean* of the posterior distributions over the *mode*; the Laplace step solely introduced to maintain recursions. The resulting method allows for fast and computationally efficient predictions; improvements can be made to make it even faster (see Section 7.2). Pivotal to the proposed approach is the formulation of the likelihood function on a continuous spatial domain.

Sequential estimation for parameters describing the dynamics of the spatiotemporal system is shown in Chapter 5. A new dual filter is developed, coined the DVBF, which allows for the incorporation of parameter uncertainty in state estimation and vice versa in an online fashion. Built on the theory of RVB, the method is seen to accrue an improvement over current state of the art methods by outperforming the DKF and being much quicker than joint state-parameter PFs. The DVBF is investigated thoroughly in the somewhat simpler temporal case, before showing to give good results in the spatiotemporal case with continuous observations. To allow for sequential estimation with point process observations, the DVBF is extended to a block estimator in an attempt to reduce the online variability in the estimates. It is seen that sequential estimation in a point process context is possible even at extremely low event counts. Naturally, with high event counts the procedure is seen to be more efficient, necessitating the consideration of a lower block size $L$.

In Chapter 6 the developed framework is applied to the Wikileaks data set, which contains tens of thousands of logs over a span of seven years. The model reduction method proposed, basis function placement and parameter estimation procedure are all seen to be easily carried out, even in this complex scenario. The efficacy of the proposed approach is evidenced by the joint field-parameter estimation procedure requiring only a few hours on a standard PC. The resulting model is then used to predict the AOG activity in 2010 based solely on data until 2009 with confidence levels; the results are encouraging and are a clear sign of the predictive capabilities gained by adopting a dynamic modelling approach to spatiotemporal point process systems. Note that an online approach could also be applied to this data set. A sequential estimator in this case may be used to identify temporal trends in the spatial growth rate which, in this case, was assumed to be constant throughout the six-year period.

## 7.2 Possible extensions

This thesis has focused on presenting an approach which may be readily applied in a variety of settings. However, one may always further relax modelling assumptions and attempt to render the procedure more computationally efficient. As a result there are several avenues for future research, two of which are discussed below.

### Application to nonlinear SPDEs

In order to render the approach developed in this thesis applicable to a more general setting, it may be extended to cater for nonlinear spatiotemporal systems and in particular nonlinear SPDEs. An example of a nonlinear SPDE is the stochastic Kuramoto-Sivashinsky equation, used in modelling surface evolution in material preparation processes [241]. In this case the Galerkin method may be readily applied, however the model reduction mechanism results in a nonlinear state-space evolution equation. In such a case alternatives to the Kalman smoother and variational Kalman smoother need to be investigated. In particular it is envisioned that the VB-Laplace approach of Chapter 4 may see considerable use in this regard; in the case of strong nonlinearities, however, it would need to be altered to cater for multimodalities [e.g. 124].

An interesting extension to the current framework is to employ a model selection approach for nonlinear systems. There has been considerable effort recently in developing methods for the selection of nonlinear terms in nonlinear autoregressive exogeneous (NARX) processes observed in noise. In principle one may use these methods [e.g. 242] for selection of nonlinear terms in the SPDE. If a VB approach (through the VB-Laplace smoother) is employed, selection between a set of candidate models may be directly employed through the computation of the lower bound $\mathcal{L}(\cdot)$ for different models [129, Section 5.3.7].

### SPDEs and GMRFs

This thesis has focused on using the SPDE explicitly for modelling spatiotemporal systems and computations were carried out using model reduced versions of the stochastic equation. A compelling way to reduce the computational burden is to proceed according to a very recent work by Lindgren et al. [20], which exploits a relationship between SPDEs and Gaussian Markov random fields (GMRFs).

A GMRF is a Gaussian random variable $\boldsymbol{x} = [x_1, \ldots, x_n]$ with Markov properties, so that for $i \neq j$, $x_i$ and $x_j$ are independent conditional on $\boldsymbol{x}^{/(i,j)}$. These Markov assumptions allow for a very sparse precision matrix $\boldsymbol{\Sigma}^{-1}$, where $\boldsymbol{\Sigma}^{-1}_{i,j} = 0$ whenever $x_i$ and $x_j$ are independent conditional on $\boldsymbol{x}^{/(i,j)}$. Such sparsity allows for considerable computational gains through the use of fast numerical algorithms for factorisations of the precision matrix [243]. GMRFs also lend themselves easily to inference using the integrated nested Laplace approximation (INLA) approach for approximate inference of the posterior marginals, a computationally efficient method employing alternating Laplace approximations of the Gaussian field and underlying parameter distributions [243]. The computational advantages gained by using GMRFs are offset by the dependence on the precision matrix; it is difficult to parameterise the matrix in order to achieve some predefined behaviour whilst ensuring positive definiteness. Nonetheless, GMRFs have been used to closely approximate covariance functions commonly used in geostatistical models [244] and have thus been used for computational gains in this regard.

The key insight in [20] is that GMRFs may be very accurately constructed explicitly from a restricted class of SPDEs which have as their solution a Gaussian random field with a Matérn covariance function, a flexible function which has the usual squared exponential function just as a special case. The authors also use a basis function expansion and employ the Galerkin method for obtaining a matrix-vector notation similar to the approach presented here. They restrict themselves to a triangulation of the spatial domain and an ensuing use of pyramidal basis functions; however the piecewise linearity of the basis functions is pivotal to obtain matrix sparsity [20, Appendix C] and the computational gains allow for an impressive amount of basis functions. On the other hand the triangulation they employ is highly inefficient when compared to the approach in this thesis, as it requires vertices at each observation location which, although a 'safe' approach, in many cases is highly inefficient. Frequency methods for basis function selection outlined in this thesis may be employed to further reduce computational cost.

Whilst the approach considered in this thesis extends to nonlinear systems with the use of adequate approximations, the approach in [20] ceases to be beneficial outside the restricted class of SPDEs considered therein. Nonetheless there is a lot to be gained by sharing insights between this work and theirs, particularly because their motivation was to use the SPDE as a *tool* for finding equivalent GMRFs for a geostatistical model; modelling using SPDEs in itself was not of prime interest. In the context of this work, it is thus worth considering equivalent GMRF models for the SPDEs (where these can be found), and attempt to obtain computational gains using this approach. Note that this would not lead to any loss in interpretability since the parameters appearing in the GMRF are directly related to the parameters appearing in the SPDE. Once inference is completed, prediction and control may then be carried out using the reduced dynamic model. It would also be interesting to apply GMRFs to spatiotemporal point process observations; as noted by Diggle in the reviewer's comments of [20], such an investigation is decidedly warranted.

## 7.3 Spatiotemporal control

It is beneficial at this stage to place this work in context, as depicted in Figure 7.1. Here three core themes are highlighted in the context of spatiotemporal systems; inference from continuous observations, inference from point process observations and control of spatiotemporal systems using static or mobile agents. Note that simultaneous consideration of all three aspects in a single framework leads to the consideration of, for instance, aggregation of continuous and point process observations in spatiotemporal inference, or the trajectory planning of mobile agents in order to simultaneously estimate and control spatiotemporal systems. This thesis has tackled two aspects shown in this figure and will now briefly propose a way forward for the third.

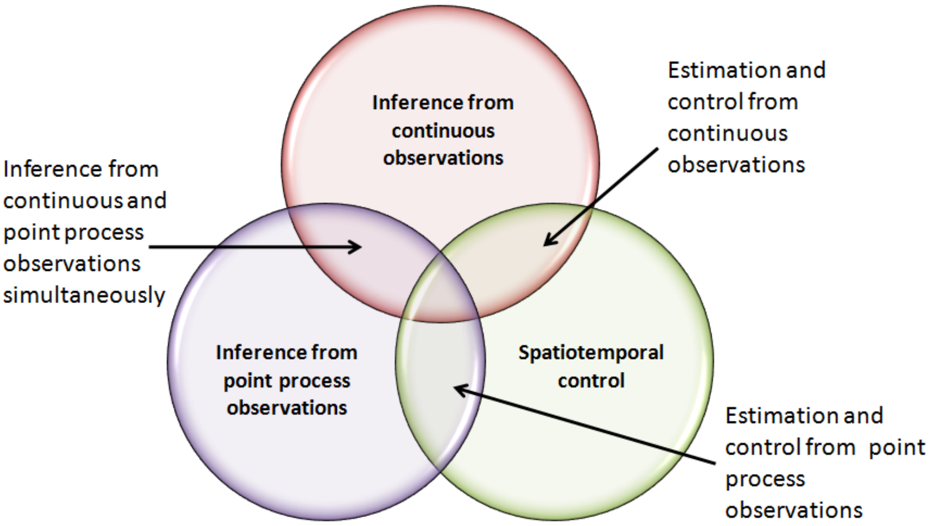

Inference from continuous observations
Estimation and control from continuous observations
Inference from continuous and point process observations simultaneously
Inference from point process observations
Spatiotemporal control
Estimation and control from point process observations

Figure 7.1: Integrated systems theoretic framework for the estimation and control of spatiotemporal systems.

### Control of spatiotemporal systems from continuous observations

Consider a spatiotemporal system deemed to be governed by a SPDE of the form (2.20) for which parameters have been estimated using any of the methods in Chapter 3 or Chapter 4. Assume now that mobile agents have been put into place to autonomously *control* the spatiotemporal field, that is, bring it to a desired reference spatial function $z_r$. The SPDE may be re-expressed as

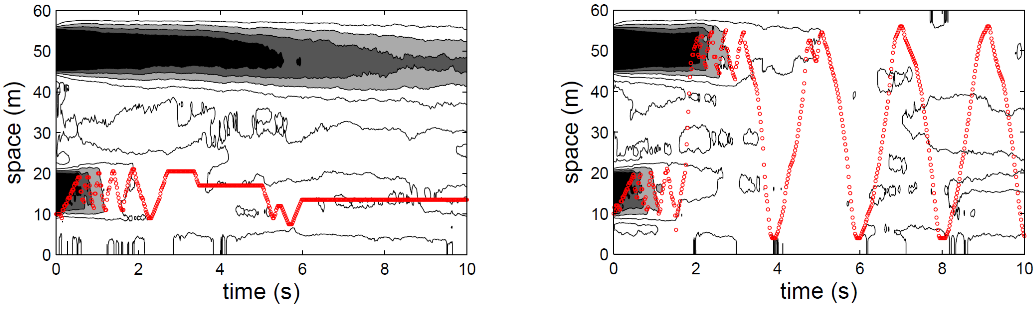
$$\begin{aligned} \mathrm{d}z(t) &= \left[\mathcal{A}z(t) + \sum_{i=1}^{m} \mathcal{B}_i(t)u_i(t)\right]\mathrm{d}t + \sigma_w \mathrm{d}W(t), \\ z(0) &= z_0 \in \mathcal{D}(\mathcal{A}), \end{aligned} \tag{7.1}$$

where the $\mathcal{B}_i(t) : \mathbb{R} \to H, i = 1 \dots m$ denotes the input operator of the $i^{th}$ agent. Each operator distributes the control action $u_i(t)$ over a finite spatial domain of compact support according to a function $b(\boldsymbol{s}; \boldsymbol{s}_i^c(t))$, usually taken to be a box function centred at the agent's location $\boldsymbol{s}_i^c(t)$.

The agents also take readings in real-time according to the usual observation equation

$$y_i(t) \quad = \quad \int_{\mathcal{O}} c(\mathbf{s}; \mathbf{s}_i^c(t)) z(t, \mathbf{s}) \mathrm{d}\mathbf{s}, \tag{7.2}$$

where $c(\mathbf{s}; \mathbf{s}_i^c(t))$ is an averaging function, which may be taken to be the same as $b(\boldsymbol{s}; \boldsymbol{s}_i^c(t))$



(a)

(b)

Figure 7.2: Trajectory optimisation (red) and control of a spatiotemporal field by (a) assuming a deterministic field and ignoring field uncertainty and (b) assuming a stochastic field with spatiotemporal varying uncertainty. In both cases the aim is to drive a strong spatiotemporal field (black areas) to zero (white).

for implementation convenience; this assumption is referred to as collocated sensing and actuation [54]. Assume the input is given according to the simple proportional control law

$$u_i(t) = -\gamma(y_i(t) - z_i^r(t)), \tag{7.3}$$

where $z_i^r(t)$ is the reference field at the $i^{th}$ agent's location. The question to ask is then, how should the agents move in order to control the spatiotemporal system in an optimal fashion? Whilst there has been considerable effort in trajectory planning for the optimal estimation of stochastic spatiotemporal systems using mobile agents [161, 245] little has been done in the context of control, and algorithms laid out for the deterministic case [54] clearly fail in the presence of random perturbations.

In [57] the author proposed a way of obtaining a very encouraging spatiotemporal control performance by incorporating uncertainty into the control performance index, in the spirit of cautious control [145] in the classical literature. The algorithm by far outperformed agents which assumed deterministic fields and which are thus discouraged to *explore* the space in a timely fashion. The performance benefit is clearly seen in Figure 7.2 where a single agent seamlessly switches between moving to control a field, and moving to explore a field; the reader is referred to [57] for further details.

Pivotal to the discussion here, the optimisation problem in [57] was tackled using the basis function modelling approach discussed extensively in this thesis (in fact it considered the same system analysed in the case study in Section 3.3). It thus could be integrated directly with modelling approaches provided in this work, giving a clear direction for future work for the third element of Figure 7.1.

### Spatiotemporal control from point process observations

Through the state-space approach considered in this thesis there is an avenue for future research in the context of spatiotemporal point process control. In this case agents would control the underlying latent process which, in turn, alters the intensity of event generation. The notion may even be entertained on a global scale; one such example is crop disease spread. Data collected from crop monitoring processes of viral diseases in staple crops [246] may be used to obtain a spatiotemporal point process model describing the dynamic spread of the disease. This model can be subsequently placed into a control framework. With appropriate input models, the spread of the disease under different control strategies may be investigated.

Another large-scale example is armed conflict. As seen in Chapter 6, it is plausible to obtain dyamic models describing the spatiotemporal spread of insurgency and other related social phenomena. Could the model, together with its predictive abilities, be used to aid decision making when deliberating peace-keeping initiatives? Could ideas from control be excercised in an attempt to stabilise an 'unstable' social phenomenon? It is still early to say; however it is believed this thesis has presented a first step in attempting to find the answers to these and other such compelling questions.

# Appendix A

# Supplementary Algorithms

This appendix contains the implementation details of the filters and smoothers developed for the study of the linear SPDEs. In particular Algorithm A.1 and Algorithm A.2 are the standard RTS smoother and variational Kalman smoother tailored for these systems. Algorithm A.3 is the novel VB-Laplace smoother required for the study of SPDEs from point process observations. Algorithm A.4 is the DVBF tailored for online estimation of SPDEs from continuous observations. Finally, Algorithm A.5 gives the modified backward pass for the linear growth model used for describing data in the AWD.

**Algorithm A.1** The Rauch Tung Striebel smoother (with cross-covariance evaluation)

```
**Input:** Data set $\mathcal{Y}, \{\boldsymbol{C}_k\}_{k=1}^K, \widetilde{\boldsymbol{Q}}, \{\boldsymbol{V}_i\}_{i=1}^d, \boldsymbol{\Psi}_{\boldsymbol{x}}, \boldsymbol{\theta}_g = (\boldsymbol{\vartheta}, \sigma_w^2, \sigma_v^2)$ and $\boldsymbol{\theta}_h = (\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0^{-1})$.
Set $\boldsymbol{A} := \boldsymbol{A}(\boldsymbol{\vartheta}), \boldsymbol{\Sigma}_w := \sigma_w^2 \widetilde{\boldsymbol{Q}}, \boldsymbol{\Sigma}_v = \sigma_v^2 \boldsymbol{I}$.

*Forward pass (Kalman filter [118])*
Set $\hat{\boldsymbol{x}}_{0|0} = \boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_{0|0} = \boldsymbol{\Sigma}_0$.
**for** $k = 1$ **to** $K$
    $\hat{\boldsymbol{x}}_{k|k-1} = \boldsymbol{A}\hat{\boldsymbol{x}}_{k-1|k-1}$
    $\boldsymbol{\Sigma}_{k|k-1} = \boldsymbol{A}\boldsymbol{\Sigma}_{k-1|k-1}\boldsymbol{A}^T + \boldsymbol{\Sigma}_w$
    $\boldsymbol{K}_k = \boldsymbol{\Sigma}_{k|k-1}\boldsymbol{C}_k^T[\boldsymbol{C}_k\boldsymbol{\Sigma}_{k|k-1}\boldsymbol{C}_k^T + \boldsymbol{\Sigma}_v]^{-1}$
    $\hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k[\boldsymbol{y}_k - \boldsymbol{C}_k\hat{\boldsymbol{x}}_{k|k-1}]$
    $\boldsymbol{\Sigma}_{k|k} = \boldsymbol{\Sigma}_{k|k-1} - \boldsymbol{K}_k\boldsymbol{C}_k\boldsymbol{\Sigma}_{k|k-1}$
**end for**

*Backward pass*
**for** $k = (K-1)$ **down to** $0$
    $\boldsymbol{G}_k = \boldsymbol{\Sigma}_{k|k}\boldsymbol{A}^T\boldsymbol{\Sigma}_{k+1|k}^{-1}$
    $\hat{\boldsymbol{x}}_{k|K} = \hat{\boldsymbol{x}}_{k|k} + \boldsymbol{G}_k(\hat{\boldsymbol{x}}_{k+1|K} - \boldsymbol{A}\hat{\boldsymbol{x}}_{k|k})$
    $\boldsymbol{\Sigma}_{k|K} = \boldsymbol{\Sigma}_{k|k} + \boldsymbol{G}_k(\boldsymbol{\Sigma}_{k+1|K} - \boldsymbol{\Sigma}_{k+1|k})\boldsymbol{G}_k^T$
**end for**

*Computation of cross-covariance* $\{\boldsymbol{M}_k\}_{k=1}^K$
$\boldsymbol{M}_{K|K} = (\boldsymbol{I}^n - \boldsymbol{K}_K\boldsymbol{C}_K)\boldsymbol{A}\boldsymbol{\Sigma}_{K-1|K-1}$
**for** $k = (K-1)$ **down to** $1$
    $\boldsymbol{M}_{k|K} = \boldsymbol{\Sigma}_{k|k}\boldsymbol{G}_{k-1}^T + \boldsymbol{G}_k(\boldsymbol{M}_{k+1|K} - \boldsymbol{A}\boldsymbol{\Sigma}_{k|k})\boldsymbol{G}_{k-1}^T$
**end for**

**Output:** $\{\hat{\boldsymbol{x}}_{k|K}, \boldsymbol{\Sigma}_{k|K}\}_{k=0}^K, \{\boldsymbol{M}_{k|K}\}_{k=1}^K$
```

**Algorithm A.2** The variational Kalman smoother (with cross-covariance evaluation)

```
Input: Data set $\mathcal{Y}$, parameters $\boldsymbol{C}$, $\widetilde{\boldsymbol{Q}}$, $\{\boldsymbol{V}_i\}_{i=1}^{d}, \boldsymbol{\Psi}_{\boldsymbol{x}}$, parameter distributions
$\tilde{p}(\boldsymbol{\theta}_g) = \tilde{p}(\boldsymbol{\vartheta})\tilde{p}(\sigma_w^2)\tilde{p}(\sigma_v^2)$ and hyperparameters $\boldsymbol{\theta}_h = (\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0^{-1})$.

Compute $\mathbb{E}[\boldsymbol{A}^T \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{A}]$ from (3.69)

Forward message
Set $\hat{\boldsymbol{x}}_{0|0} = \boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_{0|0} = \boldsymbol{\Sigma}_0$.
for  $k = 1$ to $K$
    $\boldsymbol{\Sigma}^*_{k-1} = (\boldsymbol{\Sigma}^{-1}_{k-1|k-1} + \mathbb{E}[\sigma_w^{-2}]\mathbb{E}[\boldsymbol{A}^T \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{A}])^{-1}$
    $\boldsymbol{\Sigma}_{k|k} = (\mathbb{E}[\sigma_w^{-2}]\widetilde{\boldsymbol{Q}}^{-1} + \mathbb{E}[\sigma_v^{-2}]\boldsymbol{C}^T\boldsymbol{C} - \mathbb{E}[\sigma_w^{-2}]^2\widetilde{\boldsymbol{Q}}^{-1}\mathbb{E}[\boldsymbol{A}]\boldsymbol{\Sigma}^*_{k-1}\mathbb{E}[\boldsymbol{A}^T]\widetilde{\boldsymbol{Q}}^{-1})^{-1}$
    $\hat{\boldsymbol{x}}_{k|k} = \boldsymbol{\Sigma}_{k|k}(\mathbb{E}[\sigma_v^{-2}]\boldsymbol{C}^T\boldsymbol{y}_k + \mathbb{E}[\sigma_w^{-2}]\widetilde{\boldsymbol{Q}}^{-1}\mathbb{E}[\boldsymbol{A}]\boldsymbol{\Sigma}^*_k\boldsymbol{\Sigma}^{-1}_{k-1|k-1}\hat{\boldsymbol{x}}_{k-1|k-1})$
end for

Backward message
Set $\boldsymbol{\Sigma}^{-1}_{K|K+1:K} = \boldsymbol{0}$ (ignore estimate of end condition)
for  $k = (K-1)$ down to $0$
    $\boldsymbol{\Sigma}'_{k+1} = (\boldsymbol{\Sigma}^{-1}_{k+1|k+2:K} + \mathbb{E}[\sigma_v^{-2}]\boldsymbol{C}^T\boldsymbol{C} + \mathbb{E}[\sigma_w^{-2}]\widetilde{\boldsymbol{Q}}^{-1})^{-1}$
    $\boldsymbol{\Sigma}_{k|k+1:K} = (\mathbb{E}[\sigma_w^{-2}]\mathbb{E}[\boldsymbol{A}^T\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{A}] - \mathbb{E}[\sigma_w^{-2}]^2\mathbb{E}[\boldsymbol{A}^T]\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Sigma}'_{k+1}\widetilde{\boldsymbol{Q}}^{-1}\mathbb{E}[\boldsymbol{A}])^{-1}$
    $\boldsymbol{x}_{k|k+1:K} = \mathbb{E}[\sigma_w^{-2}]\boldsymbol{\Sigma}_{k|k+1:K}\mathbb{E}[\boldsymbol{A}^T]\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Sigma}'_{k+1}(\mathbb{E}[\sigma_v^{-2}]\boldsymbol{C}^T\boldsymbol{y}_{k+1} + \boldsymbol{\Sigma}^{-1}_{k+1|k+2:K}\hat{\boldsymbol{x}}_{k+1|k+2:K})$
end for

Smoothed estimate
for  $k = 0$ to $K$
    $\boldsymbol{\Sigma}_{k|K} = (\boldsymbol{\Sigma}^{-1}_{k|k} + \boldsymbol{\Sigma}^{-1}_{k|k+1:K})^{-1}$
    $\hat{\boldsymbol{x}}_{k|K} = \boldsymbol{\Sigma}_{k|K}[\boldsymbol{\Sigma}^{-1}_{k|k}\hat{\boldsymbol{x}}_{k|k} + \boldsymbol{\Sigma}^{-1}_{k|k+1:K}\hat{\boldsymbol{x}}_{k|k+1:K}]$
end for

Computation of cross-covariance $\{\boldsymbol{M}_k\}_{k=1}^{K}$
for  $k = K$ down to $1$
    $\boldsymbol{M}_{k|K} = \mathbb{E}[\sigma_w^{-2}]\boldsymbol{\Sigma}^*_{k-1}\mathbb{E}[\boldsymbol{A}]\widetilde{\boldsymbol{Q}}^{-1}\Big[\boldsymbol{\Sigma}^{-1}_{k|k+1:K} + \mathbb{E}[\sigma_w^{-2}]\widetilde{\boldsymbol{Q}}^{-1} +$
                $\mathbb{E}[\sigma_v^{-2}]\boldsymbol{C}^T\boldsymbol{C} - \mathbb{E}[\sigma_w^{-2}]^2\widetilde{\boldsymbol{Q}}^{-1}\mathbb{E}[\boldsymbol{A}]^T\boldsymbol{\Sigma}^*_{k-1}\mathbb{E}[\boldsymbol{A}^T]\widetilde{\boldsymbol{Q}}^{-1}\Big]^{-1}$
end for

Output: $\{\hat{\boldsymbol{x}}_{k|K}, \boldsymbol{\Sigma}_{k|K}\}_{k=0}^{K}, \{\boldsymbol{M}_{k|K}\}_{k=1}^{K}$
```

**Algorithm A.3** The VB-Laplace Kalman smoother for SPDE point process systems (with cross-covariance evaluation)

```
**Input:** Data set $\mathcal{Y}$, parameters $\widetilde{\boldsymbol{Q}}, \sigma_w^2, \mu, \bar{\beta}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$ and parameter distribution $\tilde{p}(\boldsymbol{\vartheta})$.

*Forward message*
Set $\hat{\boldsymbol{x}}_{0|0} = \boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_{0|0} = \boldsymbol{\Sigma}_0$.
**for** $k = 1$ **to** $K$
    $\boldsymbol{\Sigma}^*_{k-1} = (\boldsymbol{\Sigma}^{-1}_{k-1|k-1} + \sigma_w^{-2}\mathbb{E}[\boldsymbol{A}^T\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{A}])^{-1}$
    $\tilde{\boldsymbol{\Sigma}}_k = (\sigma_w^{-2}\widetilde{\boldsymbol{Q}}^{-1} - \sigma_w^{-4}\widetilde{\boldsymbol{Q}}^{-1}\mathbb{E}[\boldsymbol{A}]\boldsymbol{\Sigma}^*_{k-1}\mathbb{E}[\boldsymbol{A}^T]\widetilde{\boldsymbol{Q}}^{-1})^{-1}$
    $\tilde{\boldsymbol{x}}_k = \sigma_w^{-2}\tilde{\boldsymbol{\Sigma}}_k\widetilde{\boldsymbol{Q}}^{-1}\mathbb{E}[\boldsymbol{A}]\boldsymbol{\Sigma}^*_{k-1}\boldsymbol{\Sigma}^{-1}_{k-1|k-1}\hat{\boldsymbol{x}}_{k-1|k-1}$
    $\hat{\boldsymbol{x}}_{k|k} = \arg\max_{\boldsymbol{x}_k} \sum_{\boldsymbol{s}_j\in\{\boldsymbol{y}_k\}} (\mu + \bar{\beta}\boldsymbol{\phi}(\boldsymbol{s}_j)^T\boldsymbol{x}_k) - \Delta_t\exp(\mu)\int_{\mathcal{O}}\exp(\bar{\beta}\boldsymbol{\phi}^T(\boldsymbol{s})\boldsymbol{x}_k)\mathrm{d}\boldsymbol{s} - \frac{1}{2}(\boldsymbol{x}_k - \tilde{\boldsymbol{x}}_k)^T\tilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}_k - \tilde{\boldsymbol{x}}_k)$
    $\boldsymbol{\Sigma}_{k|k} = \left(\tilde{\boldsymbol{\Sigma}}_k^{-1} + \Delta_t\exp(\mu)\bar{\beta}^2\int_{\mathcal{O}}\boldsymbol{\phi}(\boldsymbol{s})\boldsymbol{\phi}(\boldsymbol{s})^T\exp(\bar{\beta}\boldsymbol{\phi}(\boldsymbol{s})^T\boldsymbol{x}_{k|k})\mathrm{d}\boldsymbol{s}\right)^{-1}$
**end for**

*Backward message*
Set $\boldsymbol{\Sigma}^{-1}_{K|K+1:K} = \boldsymbol{0}$ (ignore estimate of end condition)
**for** $k = (K-1)$ **down to** $0$
    $\boldsymbol{\Sigma}'_{k+1} = \left(\boldsymbol{\Sigma}^{-1}_{k+1|k+2:K} + \Delta_t\exp(\mu)\bar{\beta}^2\int_{\mathcal{O}}\boldsymbol{\phi}(\boldsymbol{s})\boldsymbol{\phi}(\boldsymbol{s})^T\exp(\bar{\beta}\boldsymbol{\phi}(\boldsymbol{s})^T\hat{\boldsymbol{x}}_{k+1|k+1})\mathrm{d}\boldsymbol{s}\right)^{-1}$
    $\boldsymbol{x}'_{k+1} = \boldsymbol{x}_{k+1|k+1} + \boldsymbol{\Sigma}'_{k+1}\Big(\boldsymbol{\Sigma}^{-1}_{k+1|k+2:K}(\hat{\boldsymbol{x}}_{k+1|k+2:K} - \hat{\boldsymbol{x}}_{k+1|k+1})$
            $+ \sum_{\boldsymbol{s}_j\in\{\boldsymbol{y}_k\}}\bar{\beta}\boldsymbol{\phi}(\boldsymbol{s}_j) - \Delta_t\exp(\mu)\bar{\beta}\int_{\mathcal{O}}\boldsymbol{\phi}(\boldsymbol{s})\exp(\bar{\beta}\boldsymbol{\phi}(\boldsymbol{s})^T\hat{\boldsymbol{x}}_{k+1|k+1})\mathrm{d}\boldsymbol{s}\Big)$
    $\boldsymbol{\Sigma}_{k|k+1:K} = (\sigma_w^{-2}\mathbb{E}[\boldsymbol{A}^T\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{A}] - \sigma_w^{-4}\mathbb{E}[\boldsymbol{A}^T]\widetilde{\boldsymbol{Q}}^{-1}(\boldsymbol{\Sigma}'^{-1}_{k+1} + \sigma_w^{-2}\widetilde{\boldsymbol{Q}}^{-1})^{-1}\widetilde{\boldsymbol{Q}}^{-1}\mathbb{E}[\boldsymbol{A}])^{-1}$
    $\boldsymbol{x}_{k|k+1:K} = \sigma_w^{-2}\boldsymbol{\Sigma}_{k|k+1:K}\mathbb{E}[\boldsymbol{A}^T]\widetilde{\boldsymbol{Q}}^{-1}(\boldsymbol{\Sigma}'^{-1}_{k+1} + \sigma_w^{-2}\widetilde{\boldsymbol{Q}}^{-1})^{-1}\boldsymbol{\Sigma}'^{-1}_{k+1}\boldsymbol{x}'_{k+1}$
**end for**

*Smoothed estimate*
**for** $k = 0$ **to** $K$
    $\boldsymbol{\Sigma}_{k|K} = (\boldsymbol{\Sigma}^{-1}_{k|k} + \boldsymbol{\Sigma}^{-1}_{k|k+1:K})^{-1}$
    $\hat{\boldsymbol{x}}_{k|K} = \boldsymbol{\Sigma}_{k|K}[\boldsymbol{\Sigma}^{-1}_{k|k}\hat{\boldsymbol{x}}_{k|k} + \boldsymbol{\Sigma}^{-1}_{k|k+1:K}\hat{\boldsymbol{x}}_{k|k+1:K}]$
**end for**

*Computation of cross-covariance* $\{\boldsymbol{M}_k\}_{k=1}^K$
**for** $k = K$ **down to** $1$
    $\boldsymbol{M}_{k|K} = \sigma_w^{-2}\boldsymbol{\Sigma}^*_{k-1}\mathbb{E}[\boldsymbol{A}]^T\widetilde{\boldsymbol{Q}}^{-1}\Big[\boldsymbol{\Sigma}^{-1}_{k|k+1:K} + \sigma_w^{-2}\widetilde{\boldsymbol{Q}}^{-1}$
            $+\Delta_t\exp(\mu)\bar{\beta}^2\int_{\mathcal{O}}\boldsymbol{\phi}(\boldsymbol{s})\boldsymbol{\phi}(\boldsymbol{s})^T\exp(\bar{\beta}\boldsymbol{\phi}(\boldsymbol{s})^T\hat{\boldsymbol{x}}_{k|K})\mathrm{d}\boldsymbol{s} - \sigma_w^{-4}\widetilde{\boldsymbol{Q}}^{-1}\mathbb{E}[\boldsymbol{A}]\boldsymbol{\Sigma}^*_{k-1}\mathbb{E}[\boldsymbol{A}^T]\widetilde{\boldsymbol{Q}}^{-1}\Big]^{-1}$
**end for**

**Output:** $\{\hat{\boldsymbol{x}}_{k|K}, \boldsymbol{\Sigma}_{k|K}\}_{k=0}^K, \{\boldsymbol{M}_{k|K}\}_{k=1}^K$.
```

**Algorithm A.4** The DVBF for online linear SPDE dynamics estimation

**Input:** Data set $\mathcal{Y}$, parameters $\boldsymbol{C}, \widetilde{\boldsymbol{Q}}, \sigma_w^2, \sigma_v^2$ and initial parameter and state distributions $\tilde{p}(\boldsymbol{\vartheta}_0) \sim \mathcal{N}_{\boldsymbol{\vartheta}_0}(\boldsymbol{0}, \kappa_{\boldsymbol{\vartheta}}^{-1}\boldsymbol{I})$ and $\tilde{p}(\boldsymbol{x}_0) \sim \mathcal{N}_{\boldsymbol{x}_0}(\boldsymbol{0}, \kappa_{\boldsymbol{x}}^{-1}\boldsymbol{I})$ where the precisions $\kappa_{\boldsymbol{\vartheta}}$ and $\kappa_{\boldsymbol{x}}$ are small.

```
for k = 1 to $K$
    set $\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k] = \mathbb{E}_{\bar{p}(\boldsymbol{\vartheta}_{k-1})}[\boldsymbol{A}_{k-1}]$.
    set $\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k^T \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{A}_k]$
                $= \mathbb{E}_{\bar{p}(\boldsymbol{\vartheta}_{k-1})}[\boldsymbol{A}_{k-1}^T \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{A}_{k-1}]$.
    while (not converged)*

        Filter state
        $\boldsymbol{\Sigma}_{k-1}^* = (\boldsymbol{\Sigma}_{k-1|k-1}^{-1} + \sigma_w^{-2}\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k^T \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{A}_k])^{-1}$
        $\boldsymbol{\Sigma}_{k|k} = (\sigma_w^{-2}\widetilde{\boldsymbol{Q}}^{-1} + \sigma_v^{-2}\boldsymbol{C}^T\boldsymbol{C} - \sigma_w^{-4}\widetilde{\boldsymbol{Q}}^{-1}\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k]\boldsymbol{\Sigma}_{k-1}^*\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k^T]\widetilde{\boldsymbol{Q}}^{-1})^{-1}$
        $\hat{\boldsymbol{x}}_{k|k} = \boldsymbol{\Sigma}_{k|k}(\sigma_v^{-2}\boldsymbol{C}^T\boldsymbol{y}_k + \sigma_w^{-2}\widetilde{\boldsymbol{Q}}^{-1}\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k]\boldsymbol{\Sigma}_k^*\boldsymbol{\Sigma}_{k-1|k-1}^{-1}\hat{\boldsymbol{x}}_{k-1|k-1})$

        Compute smoothed statistics
        $\boldsymbol{\Sigma}' = (\sigma_v^{-2}\boldsymbol{C}^T\boldsymbol{C} + \sigma_w^{-2}\widetilde{\boldsymbol{Q}}^{-1})^{-1}$
        $\boldsymbol{\Sigma}_{k-1|k} = (\boldsymbol{\Sigma}_{k-1|k-1} + \sigma_w^{-2}\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k^T\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{A}_k]$
                $-\sigma_w^{-4}\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k^T]\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Sigma}'\widetilde{\boldsymbol{Q}}^{-1}\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k])^{-1}$.
        $\hat{\boldsymbol{x}}_{k-1|k} = \boldsymbol{\Sigma}_{k-1|k}(\sigma_w^{-2}\sigma_v^{-2}\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k^T]\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Sigma}'\boldsymbol{C}^T\boldsymbol{y}_{k+1} + \boldsymbol{\Sigma}_{k-1|k-1}^{-1}\hat{\boldsymbol{x}}_{k-1|k-1})$
        $\boldsymbol{\Lambda}_{k-1} = \boldsymbol{\Sigma}_{k-1|k-1} + \boldsymbol{x}_{k-1|k}\boldsymbol{x}_{k-1|k}^T$
        $\boldsymbol{\Gamma}_k = \sigma_w^{-2}\boldsymbol{\Sigma}_{k-1}^*\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k]\widetilde{\boldsymbol{Q}}^{-1}\Big[\sigma_w^{-2}\widetilde{\boldsymbol{Q}}^{-1} + \sigma_v^{-2}\boldsymbol{C}^T\boldsymbol{C}$
                $-\sigma_w^{-4}\widetilde{\boldsymbol{Q}}^{-1}\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k]^T\boldsymbol{\Sigma}_{k-1}^*\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k^T]\widetilde{\boldsymbol{Q}}^{-1}\Big]^{-1} + \boldsymbol{x}_{k|k}\boldsymbol{x}_{k-1|k}^T$.

        Filter parameter
        $\boldsymbol{\upsilon}_k = \Delta_t[\mathrm{tr}(\boldsymbol{V}_i^T\boldsymbol{\Psi}_{\boldsymbol{x}}^{-1}\widetilde{\boldsymbol{Q}}^{-1}(\boldsymbol{\Gamma}_k - \boldsymbol{\Lambda}_{k-1}))]_{i=1}^d$.
        $\boldsymbol{\Upsilon}_k = \Delta_t^2[\mathrm{tr}(\boldsymbol{V}_i^T\boldsymbol{\Psi}_{\boldsymbol{x}}^{-1}\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Psi}_{\boldsymbol{x}}^{-1}\boldsymbol{V}_j\boldsymbol{\Lambda}_{k-1})]_{i,j=1}^d$
        $\boldsymbol{\Sigma}_{k|k}^{\boldsymbol{\vartheta}} = \big(\sigma_w^{-2}\boldsymbol{\Upsilon}_k + \lambda\boldsymbol{\Sigma}_{k-1|k-1}^{\boldsymbol{\vartheta}^{-1}}\big)^{-1}$
        $\hat{\boldsymbol{\vartheta}}_{k|k} = \boldsymbol{\Sigma}_{k|k}^{\boldsymbol{\vartheta}}\big(\sigma_w^{-2}\boldsymbol{\upsilon}_k + \lambda\boldsymbol{\Sigma}_{k-1,k-1}^{\boldsymbol{\vartheta}^{-1}}\hat{\boldsymbol{\vartheta}}_{k-1|k-1}\big)$
        Update $\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k] = \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1}\sum_{i=1}^d \hat{\theta}_{k,i}\boldsymbol{V}_i$ and $\mathbb{E}_{\tilde{p}(\boldsymbol{\vartheta}_k)}[\boldsymbol{A}_k^T\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{A}_k]$ from (3.69).
```

**Output:** $\{\hat{\boldsymbol{x}}_{k|k}, \boldsymbol{\Sigma}_{k|k}, \hat{\boldsymbol{\vartheta}}_{k|k}, \boldsymbol{\Sigma}_{k|k}^{\boldsymbol{\vartheta}}\}_{k=1}^K$.

* Convergence is quick in the presence of relatively low observation noise. To the detriment of distributional accuracy, the loop may be run once through only, should computational time become an issue.

**Algorithm A.5** Modified backward message for the variational-Laplace Kalman smoother for point process systems with linear growth model

```
*Backward message*
Set $\boldsymbol{\Sigma}^{-1}_{K|K+1:K} = \mathbf{0}$ (ignore estimate of end condition)
**for** $k = (K-1)$ **down to** $0$
    $\boldsymbol{x}'_{k+1} = \underset{\boldsymbol{x}_{k+1}}{\arg\max} \sum_{\boldsymbol{s}_j \in \{\boldsymbol{y}_{k+1}\}} (\mu + \boldsymbol{\phi}(\boldsymbol{s}_j)^T \boldsymbol{x}_{k+1})$
        $-\Delta_t \exp(\mu) \int_{\mathcal{O}} \exp(\boldsymbol{\phi}^T(\boldsymbol{s})\boldsymbol{x}_{k+1}) \mathrm{d}\boldsymbol{s}$
        $-\frac{1}{2}(\boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}}_{k+1|k+2:K})^T \boldsymbol{\Sigma}^{-1}_{k+1|k+2:K}(\boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}}_{k+1|k+2:K})$
    $\boldsymbol{\Sigma}'_{k+1} = \left(\boldsymbol{\Sigma}^{-1}_{k+1|k+2:K} + \Delta_t \exp(\mu)\bar{\beta}^2 \int_{\mathcal{O}} \boldsymbol{\phi}(\boldsymbol{s})\boldsymbol{\phi}(\boldsymbol{s})^T \exp(\bar{\beta}\boldsymbol{\phi}(\boldsymbol{s})^T \boldsymbol{x}'_{k+1}) \mathrm{d}\boldsymbol{s}\right)^{-1}$
    $\boldsymbol{\Sigma}_{k|k+1:K} = (\sigma_w^{-2} \widetilde{\boldsymbol{Q}}^{-1} - \sigma_w^{-4} \widetilde{\boldsymbol{Q}}^{-1} (\boldsymbol{\Sigma}'^{-1}_{k+1} + \sigma_w^{-2} \widetilde{\boldsymbol{Q}}^{-1})^{-1} \widetilde{\boldsymbol{Q}}^{-1})^{-1}$
    $\boldsymbol{x}_{k|k+1:K} = \boldsymbol{\Sigma}_{k|k+1:K}(-\Delta_t \sigma_w^{-2} \widetilde{\boldsymbol{Q}}^{-1} \mathbb{E}[\boldsymbol{\theta}] + \sigma_w^{-2} \widetilde{\boldsymbol{Q}}^{-1} (\boldsymbol{\Sigma}'^{-1}_{k+1} + \sigma_w^{-2} \widetilde{\boldsymbol{Q}}^{-1})^{-1} (\boldsymbol{\Sigma}'^{-1}_{k+1} \boldsymbol{x}'_{k+1}$
        $+\Delta_t \sigma_w^{-2} \widetilde{\boldsymbol{Q}}^{-1} \mathbb{E}[\boldsymbol{\theta}]))$
**end for**
```

# Appendix B

# Update rules for EM and VBEM

## B.1 M-step for EM - Linear observation process

Throughout this section $\tilde{p}(\mathcal{X})$ denotes the distribution of the states at the $(i+1)^{th}$ iteration, i.e., that obtained by the RTS smoother of Algorithm A.1 using $\boldsymbol{\theta} = \boldsymbol{\theta}^{(i)}$. Terms which are independent from the parameter being maximised are omitted for clarity's sake.

**Finding $\boldsymbol{\mu}_0^{(i+1)}$:**

$$\begin{aligned}
\boldsymbol{\mu}_0^{(i+1)} &= \arg\max_{\boldsymbol{\mu}_0} -\frac{1}{2}\mathbb{E}_{\tilde{p}(\mathcal{X})}\left[(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)\right] \\
&= \arg\max_{\boldsymbol{\mu}_0} \hat{\boldsymbol{x}}_{0|K}^T \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}\boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 \\
&= \hat{\boldsymbol{x}}_{0|K}. \qquad \text{(B.1)}
\end{aligned}$$

**Finding ${\boldsymbol{\Sigma}_0^{-1}}^{(i+1)}$:**

$$\begin{aligned}
{\boldsymbol{\Sigma}_0^{-1}}^{(i+1)} &= \arg\max_{\boldsymbol{\Sigma}_0^{-1}} \frac{1}{2}\ln|\boldsymbol{\Sigma}_0^{-1}| - \frac{1}{2}\mathbb{E}_{\tilde{p}(\mathcal{X})}\left[(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)\right] \\
&= \arg\max_{\boldsymbol{\Sigma}_0^{-1}} \frac{1}{2}\ln|\boldsymbol{\Sigma}_0^{-1}| - \frac{1}{2}\left[\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1}\mathbb{E}_{\tilde{p}(\mathcal{X})}[\boldsymbol{x}_0\boldsymbol{x}_0^T]) - \mathrm{tr}(\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{x}}_{0|K}\hat{\boldsymbol{x}}_{0|K}^T)\right] \\
&= \arg\max_{\boldsymbol{\Sigma}_0^{-1}} \frac{1}{2}\ln|\boldsymbol{\Sigma}_0^{-1}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_{0|K}) \\
&= \boldsymbol{\Sigma}_{0|K}^{-1}, \qquad \text{(B.2)}
\end{aligned}$$

where linearity has been used to interchange the trace and expectation operators and where the maximisation has been carried out using the identity

$$\frac{\partial|\boldsymbol{\Sigma}^{-1}|}{\partial\boldsymbol{\Sigma}^{-1}} = |\boldsymbol{\Sigma}^{-1}|\boldsymbol{\Sigma}. \qquad \text{(B.3)}$$

**Finding $\boldsymbol{\vartheta}^{(i+1)}$:**

$$\boldsymbol{\vartheta}^{(i+1)} = \arg\max_{\boldsymbol{\vartheta}} -\frac{1}{2}\sum_{k=1}^{K} \sigma_w^{-2(i+1)} \mathbb{E}_{\tilde{p}(\mathcal{X})} \left[ (\boldsymbol{x}_k - \boldsymbol{A}(\boldsymbol{\vartheta})\boldsymbol{x}_{k-1})^T \widetilde{\boldsymbol{Q}}^{-1} (\boldsymbol{x}_k - \boldsymbol{A}(\boldsymbol{\vartheta})\boldsymbol{x}_{k-1}) \right]. \tag{B.4}$$

Expanding the contents within the square brackets, ignoring $\sigma_w^{-2(i+1)}$ as it is simply a scaling factor, and applying the trace operator one obtains

$$\begin{aligned}
\boldsymbol{\vartheta}^{(i+1)} &= \arg\max_{\boldsymbol{\vartheta}} \operatorname{tr}\Bigg( \boldsymbol{A}(\boldsymbol{\vartheta})^T \widetilde{\boldsymbol{Q}}^{-1} \sum_{k=1}^{K} \mathbb{E}_{\tilde{p}(\mathcal{X})}[\boldsymbol{x}_k \boldsymbol{x}_{k-1}^T] \\
&\qquad - \frac{1}{2}\boldsymbol{A}(\boldsymbol{\vartheta})^T \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{A}(\boldsymbol{\vartheta}) \sum_{k=1}^{K} \mathbb{E}_{\tilde{p}(\mathcal{X})}[\boldsymbol{x}_{k-1} \boldsymbol{x}_{k-1}^T] \Bigg) \\
&= \arg\max_{\boldsymbol{\vartheta}} \operatorname{tr}\left( \boldsymbol{A}(\boldsymbol{\vartheta})^T \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Gamma}_{1:K} - \frac{1}{2}\boldsymbol{A}(\boldsymbol{\vartheta})^T \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{A}(\boldsymbol{\vartheta}) \boldsymbol{\Lambda}_{0:K-1} \right).
\end{aligned} \tag{B.5}$$

Substituting for $\boldsymbol{A}(\boldsymbol{\vartheta})$ in (B.5) yields

$$\begin{aligned}
\boldsymbol{\vartheta}^{(i+1)} &= \arg\max_{\boldsymbol{\vartheta}} \operatorname{tr}\Bigg( [\boldsymbol{I} + \Delta_t \boldsymbol{\Psi}_{\boldsymbol{\vartheta}}^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1}] \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Gamma}_{1:K} \\
&\qquad - \frac{1}{2}[\boldsymbol{I} + \Delta_t \boldsymbol{\Psi}_{\boldsymbol{\vartheta}}^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1}]^T \widetilde{\boldsymbol{Q}}^{-1} [\boldsymbol{I} + \Delta_t \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{\vartheta}}] \boldsymbol{\Lambda}_{0:K-1} \Bigg) \\
&= \arg\max_{\boldsymbol{\vartheta}} \operatorname{tr}\Bigg( \Delta_t \boldsymbol{\Psi}_{\boldsymbol{\vartheta}}^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} [\boldsymbol{\Gamma}_{1:K} - \boldsymbol{\Lambda}_{0:K-1}] - \frac{1}{2}\Delta_t^2 \boldsymbol{\Psi}_{\boldsymbol{\vartheta}}^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{\vartheta}} \boldsymbol{\Lambda}_{0:K-1} \Bigg).
\end{aligned} \tag{B.6}$$

Recall that $\boldsymbol{\Psi}_{\boldsymbol{\vartheta}} = \sum_{i=1}^{d} \vartheta_i \boldsymbol{V}_i$ so that

$$\begin{aligned}
\boldsymbol{\vartheta}^{(i+1)} &= \arg\max_{\boldsymbol{\vartheta}} \operatorname{tr}\Bigg( \Delta_t \left[ \sum_{i=1}^{d} \vartheta_i \boldsymbol{V}_i{}^T \right] \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} [\boldsymbol{\Gamma}_{1:K} - \boldsymbol{\Lambda}_{0:K-1}] \\
&\qquad - \frac{1}{2}\Delta_t^2 \left[ \sum_{i=1}^{d} \vartheta_i \boldsymbol{V}_i{}^T \right] \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \left[ \sum_{i=1}^{d} \vartheta_i \boldsymbol{V}_i \right] \boldsymbol{\Lambda}_{0:K-1} \Bigg) \\
&= \arg\max_{\boldsymbol{\vartheta}} \Delta_t \operatorname{tr}\left( \sum_{i=1}^{d} \vartheta_i [\boldsymbol{V}_i^T \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} (\boldsymbol{\Gamma}_{1:K} - \boldsymbol{\Lambda}_{0:K-1})] \right) \\
&\qquad - \frac{1}{2}\Delta_t^2 \operatorname{tr}\left( \sum_{i=1}^{d}\sum_{j=1}^{d} \vartheta_i \boldsymbol{V}_i^T [\boldsymbol{\Psi}_{\boldsymbol{x}}^{-1} \widetilde{\boldsymbol{Q}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{x}}^{-1}] \vartheta_j \boldsymbol{V}_j \boldsymbol{\Lambda}_{0:K-1} \right),
\end{aligned} \tag{B.7}$$

to give

$$\begin{aligned}\boldsymbol{\vartheta}^{(i+1)} &= \arg\max_{\boldsymbol{\vartheta}} \boldsymbol{\vartheta}^T \boldsymbol{\upsilon} - \frac{1}{2}\boldsymbol{\vartheta}^T \boldsymbol{\Upsilon}\boldsymbol{\vartheta} \\ &= \boldsymbol{\Upsilon}^{-1}\boldsymbol{\upsilon}. \end{aligned} \tag{B.8}$$

**Finding $\sigma_w^{-2^{(i+1)}}$:**

$$\begin{aligned}\sigma_w^{-2^{(i+1)}} &= \arg\max_{\sigma_w^{-2}} \frac{Kn}{2}\ln\sigma_w^{-2} - \frac{\sigma_w^{-2}}{2}\sum_{k=1}^{K}\mathbb{E}_{\tilde{p}(\mathcal{X})}\left[(\boldsymbol{x}_k - \boldsymbol{A}(\boldsymbol{\vartheta}^{(i+1)})\boldsymbol{x}_{k-1})^T\widetilde{\boldsymbol{Q}}^{-1}(\boldsymbol{x}_k - \boldsymbol{A}(\boldsymbol{\vartheta}^{(i+1)})\boldsymbol{x}_{k-1})\right] \\ &= \arg\max_{\sigma_w^{-2}} \frac{Kn}{2}\ln\sigma_w^{-2} - \frac{\sigma_w^{-2}}{2}\mathrm{tr}\bigg(\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Lambda}_{1:K} - 2\boldsymbol{A}(\boldsymbol{\vartheta}^{(i+1)})^T\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Gamma}_{1:K} \\ &\qquad + \boldsymbol{A}(\boldsymbol{\vartheta}^{(i+1)})^T\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{A}(\boldsymbol{\vartheta}^{(i+1)})\boldsymbol{\Lambda}_{0:K-1}\bigg) \\ &= \arg\max_{\sigma_w^{-2}} \frac{Kn}{2}\ln\sigma_w^{-2} - \frac{\sigma_w^{-2}}{2}\boldsymbol{\Pi}_w \\ &= \frac{Kn}{\boldsymbol{\Pi}_w}. \end{aligned} \tag{B.9}$$

**Finding $\sigma_v^{-2^{(i+1)}}$:**

$$\begin{aligned}\sigma_v^{-2^{(i+1)}} &= \arg\max_{\sigma_v^{-2}} \frac{Km}{2}\ln\sigma_v^{-2} - \frac{\sigma_v^{-2}}{2}\sum_{k=1}^{K}\mathbb{E}_{\tilde{p}(\mathcal{X})}\left[(\boldsymbol{y}_k - \boldsymbol{C}_k\boldsymbol{x}_k)^T(\boldsymbol{y}_k - \boldsymbol{C}_k\boldsymbol{x}_k)\right] \\ &= \arg\max_{\sigma_v^{-2}} \frac{Km}{2}\ln\sigma_v^{-2} - \frac{\sigma_v^{-2}}{2}\sum_{k=1}^{K}\Big(\boldsymbol{y}_k^T\boldsymbol{y}_k - 2\boldsymbol{y}_k^T\boldsymbol{C}_k\hat{\boldsymbol{x}}_{k|K} + \mathrm{tr}(\boldsymbol{C}_k^T\boldsymbol{C}_k\boldsymbol{\Lambda}_k)\Big) \\ &= \arg\max_{\sigma_v^{-2}} \frac{Km}{2}\ln\sigma_v^{-2} - \frac{\sigma_v^{-2}}{2}\boldsymbol{\Pi}_v \\ &= \frac{Km}{\boldsymbol{\Pi}_v}. \end{aligned} \tag{B.10}$$

**Remark B.1** *In the interest of brevity, proof that the quantities are maximisers is omitted. However even by simple inspection it is easily seen that the second order partial derivatives of the functions to be maximised are all negative. For the variance terms $\boldsymbol{\Sigma}_0^{-1^{(i+1)}}, \sigma_v^{-2^{(i+1)}}$ and $\sigma_w^{-2^{(i+1)}}$ these are unconditionally negative whilst for $\boldsymbol{\mu}_0^{(i+1)}$ and $\boldsymbol{\vartheta}^{(i+1)}$ they are negative conditioned on $\boldsymbol{\Sigma}_0^{-1}$ and $\boldsymbol{\Upsilon}$ being positive definite, which follows from the definition of precision matrices and by construction from (3.39).*

## B.2 M-step for VBEM - Linear observation process

Throughout this section terms which are independent from the parameter distribution being found are omitted for clarity's sake.

**Finding** $\ln \tilde{p}(\boldsymbol{\vartheta})^{(i+1)}$**:**

$$\begin{aligned}\ln \tilde{p}(\boldsymbol{\vartheta})^{(i+1)} &= \ln p(\boldsymbol{\vartheta}) + \mathbb{E}_{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta}_g)/\boldsymbol{\vartheta}}[\ln p(\mathcal{X},\mathcal{Y},\boldsymbol{\theta})]\\ &= -\frac{1}{2}(\boldsymbol{\vartheta}-\hat{\boldsymbol{\vartheta}}_p)^T \boldsymbol{\Sigma}_{\boldsymbol{\vartheta},p}(\boldsymbol{\vartheta}-\hat{\boldsymbol{\vartheta}}_p) \qquad \text{(B.11)}\\ &\quad -\frac{1}{2}\sum_{k=1}^{K}\mathbb{E}_{\tilde{p}(\sigma_w^{-2})^{(i)}}[\sigma_w^{-2}]\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}\left[(\boldsymbol{x}_k-\boldsymbol{A}(\boldsymbol{\vartheta})\boldsymbol{x}_{k-1})^T\widetilde{\boldsymbol{Q}}^{-1}(\boldsymbol{x}_k-\boldsymbol{A}(\boldsymbol{\vartheta})\boldsymbol{x}_{k-1})\right].\end{aligned}$$

The subsequent steps follow on the lines of (B.5) - (B.8),

$$\ln \tilde{p}(\boldsymbol{\vartheta})^{(i+1)} = -\frac{1}{2}\boldsymbol{\vartheta}^T[\mathbb{E}_{\tilde{p}(\sigma_w^{-2})^{(i)}}[\sigma_w^{-2}]\,\boldsymbol{\Upsilon}+\boldsymbol{\Sigma}_{\boldsymbol{\vartheta},p}^{-1}]\boldsymbol{\vartheta}+\boldsymbol{\vartheta}^T[\mathbb{E}_{\tilde{p}(\sigma_w^{-2})^{(i)}}[\sigma_w^{-2}]\boldsymbol{\upsilon}+\boldsymbol{\Sigma}_{\boldsymbol{\vartheta},p}^{-1}\hat{\boldsymbol{\vartheta}}_p], \quad \text{(B.12)}$$

so that $\tilde{p}(\boldsymbol{\vartheta})^{(i+1)} = \mathcal{N}_{\boldsymbol{\vartheta}}(\hat{\boldsymbol{\vartheta}},\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}})$ with $\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}$ and $\hat{\boldsymbol{\vartheta}}$ as given in (3.77) and (3.78).

**Finding** $\ln \tilde{p}(\sigma_w^{-2})^{(i+1)}$**:**

$$\begin{aligned}\ln \tilde{p}(\sigma_w^{-2})^{(i+1)} &= \ln p(\sigma_w^{-2}) + \mathbb{E}_{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta}_g)/\sigma_w^{-2}}[\ln p(\mathcal{X},\mathcal{Y},\boldsymbol{\theta})]\\ &= (\alpha_{w,p}-1)\ln\sigma_w^{-2} - \beta_{w,p}\sigma_w^{-2} + \frac{Kn}{2}\ln\sigma_w^{-2}\\ &\quad -\frac{\sigma_w^{-2}}{2}\sum_{k=1}^{K}\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}\tilde{p}(\boldsymbol{\vartheta})^{(i+1)}}\left[(\boldsymbol{x}_k-\boldsymbol{A}(\boldsymbol{\vartheta})\boldsymbol{x}_{k-1})^T\widetilde{\boldsymbol{Q}}^{-1}(\boldsymbol{x}_k-\boldsymbol{A}(\boldsymbol{\vartheta})\boldsymbol{x}_{k-1})\right]\\ &= \left(\alpha_{w,p}-1+\frac{Kn}{2}\right)\ln\sigma_w^{-2} - \left(\beta_{w,p}+\frac{\boldsymbol{\Pi}'_w}{2}\right)\sigma_w^{-2}, \qquad \text{(B.13)}\end{aligned}$$

so that $\tilde{p}(\sigma_w^{-2})^{(i+1)} = \mathcal{G}a_{\sigma_w^{-2}}(\alpha_w,\beta_w)$ with $\alpha_w$ and $\beta_w$ as given in (3.79) and (3.80).

**Finding** $\ln \tilde{p}(\sigma_v^{-2})^{(i+1)}$**:**

$$\begin{aligned}\ln \tilde{p}(\sigma_v^{-2})^{(i+1)} &= \ln p(\sigma_w^{-2}) + \mathbb{E}_{\tilde{p}(\mathcal{X})\tilde{p}(\boldsymbol{\theta}_g)/\sigma_v^{-2}}[\ln p(\mathcal{X},\mathcal{Y},\boldsymbol{\theta})]\\ &= (\alpha_{v,p}-1)\ln\sigma_v^{-2} - \beta_{v,p}\sigma_v^{-2} + \frac{Km}{2}\ln\sigma_v^{-2}\\ &\quad -\frac{\sigma_v^{-2}}{2}\sum_{k=1}^{K}\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}}\left[(\boldsymbol{y}_k-\boldsymbol{C}_k\boldsymbol{x}_k)^T(\boldsymbol{y}_k-\boldsymbol{C}_k\boldsymbol{x}_k)\right]\\ &= \left(\alpha_{v,p}-1+\frac{Km}{2}\right)\ln\sigma_v^{-2} - \left(\beta_{v,p}+\frac{\boldsymbol{\Pi}_v}{2}\right)\sigma_v^{-2}, \qquad \text{(B.14)}\end{aligned}$$

so that $\tilde{p}(\sigma_v^{-2})^{(i+1)} = \mathcal{G}a_{\sigma_v^{-2}}(\alpha_v,\beta_v)$ with $\alpha_v$ and $\beta_v$ as given in (3.81) and (3.82).

The hyperparameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0^{-1}$ are updated as with conventional EM, by fitting the initial state distribution to the smoothed state at the initial time point, to give (3.84) and (3.85).

# Appendix C

# Nonparametric estimation of first-order and second-order point process statistics

If $\mathcal{P}_k$ is first-order stationary, then an estimator for $\lambda_k^{(1)}$ is given as [247, Chapter 15]

$$\lambda_k^{(1)} = \frac{N_k}{|\mathcal{O}|}. \tag{C.1}$$

In some cases, this assumption does not hold and one may instead either employ standard linear regression methods to mark out clear intensity trends [35] or employ a standard nonparametric kernel estimator [42, Section 4.3]

$$\lambda_k^{(1)}(\boldsymbol{s}) = \sum_{\boldsymbol{s}_i \in \mathcal{P}_k} \frac{k_b(||\boldsymbol{s} - \boldsymbol{s}_i||)}{c_{\mathcal{O},b}(\boldsymbol{s}_i)}. \tag{C.2}$$

Here, $c_{\mathcal{O},b}(\boldsymbol{s}_i)$ is an edge-correction factor given as $c_{\mathcal{O},b}(\boldsymbol{s}_i) = \int_{\mathcal{O}} k_b(||\boldsymbol{s} - \boldsymbol{s}_i||))\mathrm{d}\boldsymbol{s}$ and $k_b(\boldsymbol{s})$ is the *Epanečnikov kernel* which in one dimension is given as

$$k_b(s) = \frac{3}{4b}\left(1 - \frac{s^2}{b^2}\right)\mathbf{1}(|s| \leq 1). \tag{C.3}$$

A nonparametric estimator for the PACF is given by [35]

$$\hat{g}_{k,k}(\upsilon) = \frac{1}{2\pi\upsilon|\mathcal{O}|} \sum_{\boldsymbol{s}_i, \boldsymbol{s}_j \in \mathcal{P}_k}^{\neq} \frac{k_b(||\boldsymbol{s}_i - \boldsymbol{s}_j|| - \upsilon)}{\lambda_k^{(1)}(\boldsymbol{s}_i)\lambda_k^{(1)}(\boldsymbol{s}_j)w(\boldsymbol{s}_i, \boldsymbol{s}_j)}, \tag{C.4}$$

where $w(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is the fraction of the circle (in 2 dimensions) with centre $\boldsymbol{s}_i$ and radius $||\boldsymbol{s}_i - \boldsymbol{s}_j||$ lying in $\mathcal{O}$. Similarly, an estimate of the PCCF is given by

$$\hat{g}_{k,k+1}(\upsilon) = \frac{1}{2\pi\upsilon|\mathcal{O}|} \sum_{\boldsymbol{s}_i \in \mathcal{P}_k, \boldsymbol{s}_j \in \mathcal{P}_{k+1}}^{\neq} \frac{k_b(||\boldsymbol{s}_i - \boldsymbol{s}_j|| - \upsilon)}{\hat{\lambda}_k^{(1)}(\boldsymbol{s}_i)\hat{\lambda}_{k+1}^{(1)}(\boldsymbol{s}_j)w(\boldsymbol{s}_i, \boldsymbol{s}_j)}. \tag{C.5}$$

If the processes are taken to be second-order stationary also in time, to smooth out the nonparametric estimates an average over all time steps may be taken so that

$$\bar{g}_{k,k}(\upsilon) = \frac{1}{K}\sum_{k=1}^{K} \hat{g}_{k,k}(\upsilon), \qquad \bar{g}_{k,k+1}(\upsilon) = \frac{1}{K-1}\sum_{k=1}^{K-1} \hat{g}_{k,k+1}(\upsilon). \tag{C.6}$$

# Appendix D

# Online VB filter: selected derivations

## D.1 Computation of $\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\boldsymbol{A}_k^T \boldsymbol{\Sigma}_w^{-1} \boldsymbol{A}_k]$ for linear state-space models

Let $\boldsymbol{\Sigma}_w^{-1} = [\bar{q}_{i,j}]_{i,j=1}^n$ and consider the following standard result [248, Chapter 2]

**Theorem D.1** *For all $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{n \times n}$, $vec(\boldsymbol{ABC}) = (\boldsymbol{C}^T \otimes \boldsymbol{A}) vec(\boldsymbol{B})$*

To manage this quantity, start off by considering the case with $n = 2$. By application of Theorem D.1, the vectorization of the product of $\boldsymbol{A}_k^T \boldsymbol{\Sigma}_w^{-1} \boldsymbol{A}_k$ becomes

$$\begin{aligned} vec(\boldsymbol{A}_k^T \boldsymbol{\Sigma}_w^{-1} \boldsymbol{A}_k) &= (\boldsymbol{A}_k^T \otimes \boldsymbol{A}_k^T) vec(\boldsymbol{\Sigma}_w^{-1}) \\ &= (\boldsymbol{A}_k \otimes \boldsymbol{A}_k)^T vec(\boldsymbol{\Sigma}_w^{-1}) \\ &= \left[\begin{array}{cc|cc} a_{1,1}^2 & a_{1,1}a_{1,2} & a_{1,2}a_{1,1} & a_{1,2}^2 \\ a_{1,1}a_{2,1} & a_{1,1}a_{2,2} & a_{1,2}a_{2,1} & a_{1,2}a_{2,2} \\ \hline a_{2,1}a_{1,1} & a_{2,1}a_{1,2} & a_{2,2}a_{1,1} & a_{2,2}a_{1,2} \\ a_{2,1}^2 & a_{2,1}a_{2,2} & a_{2,2}a_{2,1} & a_{2,2}^2 \end{array}\right]^T \begin{bmatrix} \bar{q}_{1,1} \\ \bar{q}_{2,1} \\ \bar{q}_{1,2} \\ \bar{q}_{2,2} \end{bmatrix} \\ &= \boldsymbol{A}_k^{*^T} vec(\boldsymbol{\Sigma}_w^{-1}), \end{aligned} \tag{D.1}$$

where $\boldsymbol{A}_k^* = (\boldsymbol{A}_k \otimes \boldsymbol{A}_k)$. The solid lines indicate partitions which can be useful in finding a closed form for $\boldsymbol{A}_k^*$. First define a *shaping* operator which reshapes a matrix or a vector into any desired (eligible) shape. In particular define $pack_{s_2}^{s_1}(\cdot)$ as an operator which arranges matrix partitions of size $s_1$ into a square matrix of partition size $s_2$. For instance for a matrix $\boldsymbol{B} \in \mathbb{R}^{n^2\times 1}$, if $s_1 = 1 \times 1$ and $s_2 = n \times n$, the packing operator simply reverses the vectorization process. However, for $\boldsymbol{B} \in \mathbb{R}^{n^3\times n}$, $s_1 = n \times n$ and $s_2 = n \times n$, the packing operator takes matrix partitions of size $n \times n$ and reshapes them into a square matrix of final size $n^2 \times n^2$. One can rewrite (D.1)

$$\boldsymbol{A}_k^* = \left[\begin{array}{c|c} pack_{2\times 2}^{1\times 1}([vec(\boldsymbol{A}_k)vec(\boldsymbol{A}_k)^T]_{:,1}) & pack_{2\times 2}^{1\times 1}([vec(\boldsymbol{A}_k)vec(\boldsymbol{A}_k)^T]_{:,3}) \\ \hline pack_{2\times 2}^{1\times 1}([vec(\boldsymbol{A}_k)vec(\boldsymbol{A}_k)^T]_{:,2}) & pack_{2\times 2}^{1\times 1}([vec(\boldsymbol{A}_k)vec(\boldsymbol{A}_k)^T]_{:,4}) \end{array}\right]. \tag{D.2}$$

where the subscript $:, i$ is used to denote the $i^{th}$ column of the matrix. For arbitrary $n$ one can hence concisely express the expectation of $\boldsymbol{A}_k^*$ as

$$\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\boldsymbol{A}_k^*] = pack_{n\times n}^{n\times n}\Big(\{pack_{n\times n}^{1\times 1}([\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[vec(\boldsymbol{A}_k)vec(\boldsymbol{A}_k)^T]_{:,i})\}_{i=1}^n\Big). \tag{D.3}$$

This formulation, albeit tedious, can be programmed in two lines of code. Moreover, the expectations are readily computed since, for $\boldsymbol{\theta}_k = vec(\boldsymbol{A}_k)$

$$\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\boldsymbol{\theta}_k\boldsymbol{\theta}_k^T] = \boldsymbol{\Sigma}_{k|k}^{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}}_{k|k}\hat{\boldsymbol{\theta}}_{k|k}^T. \tag{D.4}$$

One hence has that

$$vec(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\boldsymbol{A}_k^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{A}_k]) = (\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_k)}[\boldsymbol{A}_k^*])^T vec(\boldsymbol{\Sigma}_w^{-1}), \tag{D.5}$$

and one final repacking with the operator $pack_{n\times n}^{1\times 1}(\cdot)$ is required to change the vector back into a matrix of appropriate size.

## D.2 Online variational posterior $\tilde{p}(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})$ for linear state-space models

The parameter variational posterior is given by

$$\tilde{p}(\boldsymbol{\theta}_k) \propto \tilde{p}(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})\exp\Bigg(-\mathbb{E}_{\tilde{p}(\mathcal{X}_k)}\left[\frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{A}_k\boldsymbol{x}_{k-1})^T\boldsymbol{\Sigma}_w^{-1}(\boldsymbol{x}_k - \boldsymbol{A}_k\boldsymbol{x}_{k-1})\right]\Bigg), \tag{D.6}$$

where $\tilde{p}(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}) = \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\theta}_{k-1})}[\ln p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})])$. Proceed by noting that the exponent is a scalar and that the trace operator can be hence applied. Excluding terms which are not a function of $\boldsymbol{\theta}_k$ and by the rotational property of the trace operator

$$\tilde{p}(\boldsymbol{\theta}_k) \propto \tilde{p}(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}) \exp(tr(\boldsymbol{\Gamma}_k \boldsymbol{A}_k^T \boldsymbol{\Sigma}_w^{-1} - \boldsymbol{\Lambda}_{k-1} \boldsymbol{A}_k^T \boldsymbol{\Sigma}_w^{-1} \boldsymbol{A}_k)). \tag{D.7}$$

Consider now the following standard result [248, Chapter 2]:

**Theorem D.2** *For all $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{n\times n}$, $tr(\boldsymbol{A}^T\boldsymbol{B}\boldsymbol{C}\boldsymbol{D}^T) = vec(\boldsymbol{A})^T(\boldsymbol{D}\otimes \boldsymbol{B})vec(\boldsymbol{C})$.*

Applying Theorem D.2 to (D.7)

$$tr(\boldsymbol{A}_k^T \boldsymbol{\Sigma}_w^{-1} \boldsymbol{A}_k \boldsymbol{\Lambda}_{k-1}) = vec(\boldsymbol{A}_k)^T [\boldsymbol{\Lambda}_{k-1} \otimes \boldsymbol{\Sigma}_w^{-1}] vec(\boldsymbol{A}_k), \tag{D.8}$$

$$tr(\boldsymbol{A}_k^T \boldsymbol{\Sigma}_w^{-1} \boldsymbol{I} \boldsymbol{\Gamma}_k) = vec(\boldsymbol{A}_k)^T [\boldsymbol{\Gamma}_k^T \otimes \boldsymbol{\Sigma}_w^{-1}] vec(\boldsymbol{I}), \tag{D.9}$$

Since

$$\tilde{p}(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}) = \mathcal{N}_{\boldsymbol{\theta}_k}(\hat{\boldsymbol{\theta}}_{k-1|k-1}, \lambda^{-1} \boldsymbol{\Sigma}^{\boldsymbol{\theta}}_{k-1|k-1}),$$

one has that

$$\begin{aligned}\tilde{p}(\boldsymbol{\theta}_k) \propto \exp\Bigg( &-\frac{1}{2}\boldsymbol{\theta}_k^T(\lambda \boldsymbol{\Sigma}^{\boldsymbol{\theta}^{-1}}_{k-1|k-1} + \boldsymbol{\Lambda}_{k-1} \otimes \boldsymbol{\Sigma}_w^{-1})\boldsymbol{\theta}_k \\ &+ \boldsymbol{\theta}_k^T[\lambda \boldsymbol{\Sigma}^{\boldsymbol{\theta}^{-1}}_{k-1|k-1} \hat{\boldsymbol{\theta}}_{k-1|k-1} + [\boldsymbol{\Gamma}_k^T \otimes \boldsymbol{\Sigma}_w^{-1}] vec(\boldsymbol{I})]\Bigg),\end{aligned} \tag{D.10}$$

which is normally distributed with mean $\hat{\boldsymbol{\theta}}_{k|k}$ and covariance $\boldsymbol{\Sigma}^{\boldsymbol{\theta}}_{k|k}$ as given in the main text.

# Bibliography

[1] M. Jefferson, “A critical note on the efficacy of the UK’s renewable obligation system as it applies to onshore wind energy developments in England,” *London Metropolitan University, Centre for International Business and Sustainability Working Papers*, vol. 1, 2008.

[2] J. Haslett and A. E. Raftery, “Space-time modelling with long-memory dependence: assessing Ireland’s wind power resource,” *Journal of the Royal Statistical Society C*, vol. 38, no. 1, pp. 1–50, 1989.

[3] M. Abeles, H. Bergman, E. Margalit, and E. Vaadia, “Spatiotemporal firing patterns in the frontal cortex of behaving monkeys,” *Journal of Neurophysiology*, vol. 70, no. 4, pp. 1629–1638, 1993.

[4] H. Kang *et al.*, “Spatio-temporal transcriptome of the human brain,” *Nature*, vol. 478, no. 7370, pp. 483–489, 2011.

[5] M. P. Ward, D. Maftei, C. Apostu, and A. Suru, “Geostatistical visualisation and spatial statistics for evaluation of the dispersion of epidemic highly pathogenic avian influenza subtype H5N1,” *Veterinary Research*, vol. 39, no. 3, pp. 22–22, 2008.

[6] T. A. Abeku *et al.*, “Malaria epidemic early warning and detection in African highlands,” *Trends in Parasitology*, vol. 20, no. 9, pp. 400–205, 2004.

[7] G. King, “Ensuring the data-rich future of the social sciences,” *Science*, vol. 331, no. 6018, pp. 719–721, 2011.

[8] N. Johnson *et al.*, “Pattern in escalations in insurgent and terrorist activity,” *Science*, vol. 333, pp. 81–84, 2011.

[9] L. Quéré *et al.*, “Trends in the sources and sinks of carbon dioxide,” *Nature Geoscience*, vol. 2, no. 12, pp. 831–836, 2009.

[10] R. G. Baraniuk, “More is less: signal processing and the data deluge,” *Science*, vol. 331, no. 6018, p. 717, 2011.

[11] L. Bachelier, *Théorie de la Speculation.* Paris: Gauthier-Villar, 1900.

[12] A. Fassò, M. Cameletti, and O. Nicolis, “Air quality monitoring using heterogeneous networks,” *Environmetrics*, vol. 18, no. 3, pp. 245–264, 2007.

[13] T. E. Unny, “Stochastic partial differential equations in groundwater hydrology,” *Stochastic Hydrology and Hydraulics*, vol. 3, no. 2, pp. 135–153, 1989.

[14] V. Šmídl and A. Quinn, *The Variational Bayes Method in Signal Processing.* New York: Springer-Verlag, 2005.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society B*, vol. 39, no. 1, pp. 1–38, 1977.

[16] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods.* New York: Springer-Verlag, 2004.

[17] P. Guttorp and P. D. Sampson, “Methods for estimating heterogeneous spatial covariance functions with environmental applications,” in *Handbook of statistics, volume 12*, G. P. Patil and C. R. Rao, Eds. Amsterdam: Elsevier Science, 1994, pp. 661–689.

[18] T. Gneiting, M. G. Genton, and P. Guttorp, “Geostatistical space-time models, stationarity, separability and full symmetry,” in *Statistics of Spatio-Temporal Systems. Monographs in Statistics and Applied Probability*, B. Finkenstaedt, L. Held, and V. Isham, Eds. Boca Raton, Florida: Chapman & Hall/CRC Press, 2007, pp. 151–175.

[19] J. A. Duan, A. E. Gelfand, and C. F. Sirmans, “Modeling space-time data using stochastic differential equations,” *Bayesian Analysis*, vol. 4, no. 4, pp. 733–758, 2009.

[20] F. Lindgren, H. Rue, and J. Lindström, “An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach,” *Journal of the Royal Statistical Society B*, vol. 73, no. 4, pp. 423–498, Sep. 2011.

[21] G. Storvik, A. Frigessi, and D. Hirst, “Stationary space-time Gaussian fields and

their time autoregressive representation," *Statistical Modelling*, vol. 2, no. 2, p. 139, 2002.

[22] J. R. Stroud, P. Müller, and B. Sanso, "Dynamic models for spatiotemporal data," *Journal of the Royal Statistical Society B*, vol. 63, pp. 673–689, 2001.

[23] M. Dewar, K. Scerri, and V. Kadirkamanathan, "Data-driven spatio-temporal modeling using the integro-difference equation," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 83–91, Jan. 2009.

[24] N. E. Leonard *et al.*, "Collective motion, sensor networks, and ocean sampling," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 48–74, 2007.

[25] S. Gibson and B. Ninness, "Robust maximum-likelihood estimation of multivariable dynamic systems," *Automatica*, vol. 41, no. 10, pp. 1667–1682, 2005.

[26] D. R. Cox and V. Isham, *Point Processes.* New York: Chapman and Hall, 1980.

[27] M. B. Hooten and C. K. Wikle, "A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove," *Environmental and Ecological Statistics*, vol. 15, no. 1, pp. 59–70, Mar. 2008.

[28] R. Barbieri, E. C. Matten, A. A. Alabi, and E. N. Brown, "A point-process model of human heartbeat intervals: new definitions of heart rate and heart rate variability," *AJP – Heart and Circulatory Physiology*, vol. 288, no. 1, pp. 424–435, Jan. 2005.

[29] A. C. Smith and E. N. Brown, "Estimating a state-space model from point process observations," *Neural Computation*, vol. 15, no. 5, pp. 965–991, 2003.

[30] A. Ergün, R. Barbieri, U. T. Eden, M. A. Wilson, and E. N. Brown, "Construction of point process adaptive filter algorithms for neural systems using sequential Monte Carlo methods," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 3, pp. 419–428, Mar. 2007.

[31] U. T. Eden, L. M. Frank, R. Barbieri, V. Solo, and E. N. Brown, "Dynamic analysis of neural encoding by point process adaptive filtering," *Neural Computation*, vol. 16, no. 5, pp. 971–998, 2004.

[32] R. P. Adams, I. Murray, and D. J. C. MacKay, "Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 9–16.

[33] F. P. Schoenberg, D. R. Brillinger, and P. Guttorp, “Point processes, spatial-temporal,” in *Encyclopedia of Environmetrics*, A. H. El-Shaarawi and W. W. Piegorsch, Eds. Chichester: John Wiley & Sons, Ltd, 2006, pp. 1573—1577.

[34] Y. Ogata, “Space-time point-process models for earthquake occurrences,” *Annals of the Institute of Statistical Mathematics*, vol. 50, no. 2, pp. 379–402, 1998.

[35] A. Brix and J. Møller, “Space-time multi type log Gaussian Cox processes with a view to modelling weeds,” *Scandinavian Journal of Statistics*, vol. 28, no. 3, pp. 471–488, 2001.

[36] P. J. Diggle, “Spatio-temporal point processes: methods and applications,” *Johns Hopkins University, Dept. of Biostatistics Working Papers*, vol. 78, 2005.

[37] A. Brix and P. J. Diggle, “Spatiotemporal prediction for log-Gaussian Cox processes,” *Journal of the Royal Statistical Society B*, vol. 63, no. 4, pp. 823–841, 2001.

[38] R. P. S. Mahler, “Multitarget Bayes filtering via first-order multitarget moments,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1152–1178, 2003.

[39] B. T. Vo, B. N. Vo, and A. Cantoni, “Analytic implementations of the cardinalized probability hypothesis density filter,” *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3553–3567, 2007.

[40] D. Clark, A. T. Cemgil, P. Peeling, and S. Godsill, “Multi-object tracking of sinusoidal components in audio with the gaussian mixture probability hypothesis density filter,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 339–342.

[41] P. C. Kyriakidis and A. G. Journel, “Geostatistical space-time models: a review.” *Journal of Mathematical Geology*, vol. 31, pp. 651–684, 1999.

[42] J. Møller and R. P. Waagepetersen, *Statistical Inference and Simulation for Spatial Point Processes.* Boca Raton, Florida: CRC Press, 2004.

[43] P. Diggle, B. Rowlingson, and T. Su, “Point process methodology for on-line spatio-temporal disease surveillance,” *Environmetrics*, vol. 16, no. 5, pp. 423–434, 2005.

[44] D. Uciński, “Optimal sensor location for parameter estimation of distributed processes,” *International Journal of Control*, vol. 73, pp. 1235–1248, 2000.

[45] H. T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems.* Boston: Birkhäser, 1989.

[46] D. Coca and S. A. Billings, "Direct parameter identification of distributed parameter systems," *International Journal of Systems Science*, vol. 31, pp. 11–17, Jun. 2000.

[47] K. Xu, C. K. Wikle, and N. I. Fox, "A kernel-based spatio-temporal dynamical model for nowcasting radar precipitation," *Journal of the American Statistical Association*, vol. 100, no. 472, pp. 1133–1144, 2005.

[48] D. Coca and S. A. Billings, "Analysis and reconstruction of stochastic coupled map lattice models," *Physics Letters A*, vol. 315, no. 1-2, pp. 61–75, 2003.

[49] C. K. Wikle and M. B. Hooten, "Hierarchical Bayesian spatio-temporal models for population spread," in *Applications of Computational Statistics in the Environmental Sciences: Hierarchical Bayes and MCMC Methods*, J. S. Clark and A. Gelfand, Eds. Oxford: Oxford University Press, 2006, pp. 145–169.

[50] C. K. Wikle, "A kernel-based spectral model for non-Gaussian spatio-temporal processes," *Statistical Modelling*, vol. 2, no. 1, pp. 299–314, 2002.

[51] M. A. Demetriou, A. Paskaleva, O. Vayena, and H. Doumanidis, "Scanning actuator guidance scheme in a 1-D thermal manufacturing process," *IEEE Transactions on Control Systems Technology*, vol. 11, no. 5, pp. 757–764, 2003.

[52] O. V. Iftime and M. A. Demetriou, "Optimal control of switched distributed parameter systems with spatially scheduled actuators," *Automatica*, vol. 45, no. 2, pp. 312–323, 2009.

[53] M. A. Demetriou, "Guidance of a moving collocated actuator/sensor for improved control of distributed parameter systems," in *IEEE Conference on Decision and Control, 2008*, 2008, pp. 215–220.

[54] ——, "Guidance of mobile actuator-plus-sensor networks for improved control and estimation of distributed parameter systems," *IEEE Transactions on Automatic Control*, vol. 55, no. 7, pp. 1570–1584, 2010.

[55] A. Zammit Mangion, K. Yuan, V. Kadirkamanathan, M. Niranjan, and G. Sanguinetti, "Online variational inference for state-space models with point process observations," *Neural Computation*, vol. 23, no. 8, pp. 1967–1999, Aug. 2011.

[56] A. Zammit Mangion, G. Sanguinetti, and V. Kadirkamanathan, "A variational approach for the online dual estimation of spatiotemporal systems governed by the IDE," in *Proceedings of the 18th IFAC World Congress*, 2011, pp. 3204–3209.

[57] A. Zammit Mangion, S. Anderson, and V. Kadirkamanathan, "Exploration and control of stochastic spatiotemporal systems with mobile agents," in *Proceedings of the 18th IFAC World Congress*, 2011, pp. 4489–4494.

[58] Y. Pan and S. A. Billings, "The identification of complex spatiotemporal patterns using coupled map lattice models," *International Journal of Bifurcation and Chaos*, vol. 18, no. 4, pp. 997–1013, 2008.

[59] S. A. Billings and D. Coca, "Identification of coupled map lattice models of deterministic distributed parameter systems," *International Journal of Systems Science*, vol. 33, no. 8, pp. 623–634, 2002.

[60] N. Parekh, S. Parthasarathy, and S. Sinha, "Global and local control of spatiotemporal chaos in coupled map lattices," *Physical Review Letters*, vol. 81, pp. 1401–1404, Aug. 1998.

[61] L. A. Bunimovich, "Coupled map lattices: one step forward and two steps back," *Physica D: Nonlinear Phenomena*, vol. 86, no. 1-2, pp. 248–255, 1995.

[62] K. Kaneko, "Overview of coupled map lattices," *Chaos*, vol. 2, no. 3, pp. 279–282, 1992.

[63] ——, "Pattern dynamics in spatiotemporal chaos: pattern selection, diffusion of defect and pattern competition intermettency," *Physica D: Nonlinear Phenomena*, vol. 34, no. 1-2, pp. 1–41, 1989.

[64] H. Richter, "Coupled map lattices as spatio-temporal fitness functions: landscape measures and evolutionary optimization," *Physica D: Nonlinear Phenomena*, vol. 237, no. 2, pp. 167–186, 2008.

[65] J. Jost, "Spectral properties and synchronization in coupled map lattices," *Physical Review E*, vol. 65, no. 1, pp. 016 201.1–016 201.9, 2001.

[66] T. Yanagita, "Phenomenology of boiling: a coupled map lattice model," *Chaos*, vol. 2, p. 343, 1992.

[67] T. Yanagita and K. Kaneko, "Modeling and characterization of cloud dynamics," *Physical Review Letters*, vol. 78, no. 22, pp. 4297–4300, 1997.

[68] M. Shen, G. Chang, S. Wang, and P. Beadle, “Nonlinear dynamics of EEG signal based on coupled network lattice model,” *Advances in Neural Networks-ISNN 2006*, pp. 560–565, 2006.

[69] D. Coca and S. A. Billings, “Identification of coupled map lattice models of complex spatio-temporal patterns,” *Physics Letters*, vol. A287, no. 1-2, pp. 65–73, Aug. 2001.

[70] M. Kot, M. A. Lewis, and P. van den Driessche, “Dispersal data and the spread of invading organisms,” *Ecology*, vol. 77, no. 7, pp. 2027–2042, Oct. 1996.

[71] M. Kot and W. M. Schaffer, “Discrete-time growth-dispersal models,” *Mathematical Biosciences*, vol. 80, no. 1, pp. 109–136, 1986.

[72] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning.* Cambridge, MA: The MIT Press, 2006.

[73] K. Scerri, M. Dewar, and V. Kadirkamanathan, “Estimation and model selection for an IDE-based spatio-temporal model,” *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 482–492, Feb. 2009.

[74] D. R. Freestone *et al.*, “A data-driven framework for neural field modeling,” *NeuroImage*, vol. 56, no. 3, pp. 1043–1058, Jun. 2011.

[75] M. B. Hooten, C. K. Wikle, R. M. Dorazio, and J. A. Royle, “Hierarchical spatiotemporal matrix models for characterizing invasions,” *Biometrics*, vol. 63, no. 2, pp. 558–567, 2007.

[76] K. Scerri, “A Systems Approach to Spatio-Temporal Modelling,” Ph.D. dissertation, University of Sheffield, 2010.

[77] M. P. Coleman, *An Introduction to Partial Differential Equations with Matlab.* London: Chapman and Hall/CRC, 2005.

[78] J. M. Hill, *Heat Conduction.* Oxford: Blackwell Scientific, 1987.

[79] M. I. Asensio and L. Ferragut, “On a wildland fire model with radiation,” *International Journal for Numerical Methods in Engineering*, vol. 54, no. 1, pp. 137–157, 2002.

[80] E. E. Holmes, M. A. Lewis, J. E. Banks, and R. R. Veit, “Partial differential equations in ecology: spatial interactions and population dynamics,” *Ecology*, vol. 75, no. 1, pp. 17–29, 1994.

[81] A. F. Bennett, *Inverse Modeling of the Ocean and Atmosphere.* Cambridge, UK: Cambridge University Press, 2002.

[82] L. C. Evans, *Partial Differential Equations (Graduate Studies in Mathematics, Vol. 19).* Providence, RI: AMS, 1998.

[83] R. C. Dalang and N. E. Frangos, “The stochastic wave equation in two spatial dimensions,” *Annals of Probability*, vol. 26, no. 1, pp. 187–212, 1998.

[84] R. A. Carmona, *Stochastic Partial Differential Equations: Six Perspectives.* American Mathematical Society, 1998.

[85] C. Prévôt and M. Röckner, *A Concise Course on Stochastic Partial Differential Equations.* Berlin: Springer-Verlag, 2007.

[86] G. Grün, K. Mecke, and M. Rauscher, “Thin-film flow influenced by thermal noise,” *Journal of Statistical Physics*, vol. 122, no. 6, pp. 1261–1291, 2006.

[87] J. B. Walsh, “A stochastic model of neural response,” *Advances in Applied Probability*, vol. 13, no. 2, pp. 231–281, Jun. 1981.

[88] G. da Prato, A. Debussche, and R. Temam, “Stochastic Burgers’ equation,” *Nonlinear Differential Equations and Applications*, vol. 1, no. 4, pp. 389–402, Dec. 1994.

[89] H. Krim and Y. Bao, “A stochastic diffusion approach to signal denoising,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 4, 1999, pp. 1773–1776.

[90] J. Duan and B. Goldys, “Ergodicity of stochastically forced large scale geophysical flows,” *International Journal of Mathematics and Mathematical Sciences*, vol. 28, no. 6, pp. 313–320, 2001.

[91] D. Coca and S. A. Billings, “Identification of finite dimensional models of infinite dimensional dynamical systems,” *Automatica*, vol. 38, no. 11, pp. 1851–1865, 2002.

[92] L. Guo and S. A. Billings, “Identification of partial differential equation models for continuous spatio-temporal dynamical systems,” *IEEE Transactions on Circuits and Systems II Express Briefs*, vol. 53, no. 8, pp. 657–661, Aug. 2006.

[93] V. Solo, “Identification of a noisy stochastic heat equation with the EM algorithm,” in *Proceedings of the 41st IEEE Conference on Decision and Control*, vol. 4, 2002, pp. 4505–4508.

[94] C. Grossmann, H. G. Roos, and M. Stynes, *Numerical Treatment of Partial Differential Equations.* Berlin: Springer-Verlag, 2007.

[95] A. M. Davie and J. G. Gaines, "Convergence of numerical schemes for the solution of parabolic stochastic partial differential equations," *Mathematics of Computation*, vol. 70, no. 233, pp. 121–134, Feb. 2000.

[96] T. E. Duncan, B. Pasik-Duncan, and P. Zimmer, "Computational methods for some stochastic partial differential equations," in *Proceedings of the 35th IEEE Conference on Decision and Control*, 1996.

[97] T. Shardlow, "Numerical methods for stochastic parabolic PDEs," *Numerical Functional Analysis and Optimization*, vol. 20, no. 1, pp. 121–145, 1999.

[98] E. Hausenblas, "Approximation for semilinear stochastic evolution equations," *Potential Analysis*, vol. 18, no. 2, pp. 141–186, 2003.

[99] R. F. Harrington, *Field Computation by Moments Method.* Piscataway: IEEE Press, 1993.

[100] B. B. King and D. A. Krueger, "The 1-D convection diffusion equation: Galerkin least squares approximations and feedback control," in *Proceedings of the IEEE Conference on Decision and Control*, vol. 2, 2004, pp. 1502–1507.

[101] W. Grecksch and P. E. Kloeden, "Time-discretised Galerkin approximations of parabolic stochastic PDEs," *Bulletin of the Australian Mathematical Society*, vol. 54, no. 01, pp. 79–85, 1996.

[102] A. Lang, "Simulation of stochastic partial differential equations and stochastic active contours," Ph.D. dissertation, Universität Mannheim, 2007.

[103] Y. Yan, "Semidiscrete Galerkin approximation for a linear stochastic parabolic partial differential equation driven by an additive noise," *BIT Numerical Mathematics*, vol. 44, no. 4, pp. 829–847, Dec. 2004.

[104] A. H. Jazwinski, *Stochastic Processes and Filtering Theory.* London: Academic Press, 1970.

[105] J. R. Leigh, *Functional Analysis and Linear Control Theory.* London: Academic Press, 1980.

[106] G. da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions.* Cambridge, UK: Cambridge University Press, 1993.

[107] J. B. Walsh, *An introduction to stochastic partial differential equations*, ser. Lecture Notes in Mathematics. Berlin: Springer, 1986, vol. 1180, pp. 265–439.

[108] K. J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations.* New York: Springer-Verlag, 2000.

[109] B. Simon, *Trace Ideals and their Applications*, 2nd ed. American Mathematical Society, 2005.

[110] J. M. Steele, *The Cauchy-Schwarz Master Class.* Cambridge, UK: Cambridge University Press, 2004.

[111] A. Abuding, "Finite element simulations of stochastic partial differential equations," Master's thesis, Chalmers University of Technology, 2007.

[112] J. Roubal, V. Havlena, and M. Benes, "Base vectors for solving partial differential equations," in *Proceedings of the International Control Conference*, 2006.

[113] M. Dewar and V. Kadirkamanathan, "A canonical space-time state space model: state and parameter estimation," *IEEE Transactions on Signal Processing*, vol. 55, no. 10, pp. 4862–4870, Oct. 2007.

[114] F. Sawo, V. Klumpp, and U. D. Hanebeck, "Simultaneous state and parameter estimation of distributed-parameter physical systems based on sliced Gaussian mixture filter," in *Proceedings of the 11th International Conference on Information Fusion*, 2008, pp. 1–8.

[115] M. Briers, A. Doucet, and S. Maskell, "Smoothing algorithms for state-space models," University of Cambridge, Department of Engineering, Tech. Rep. TR-CUED-F-INFENG 498, 2004.

[116] P. Ahrendt, "The multivariate Gaussian probability distribution," Technical University of Denmark, Tech. Rep. IMM2005-03312, 2005.

[117] Y. Bar-Shalom, X. R. Li, T. Kirubarajan, and J. Wiley, *Estimation with Applications to Tracking and Navigation.* New York: John Wiley & Sons, Inc., 2001.

[118] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.

[119] E. A. Wan and A. T. Nelson, "Dual extended Kalman filter methods," in *Kalman Filtering and Neural Networks*, S. Haykin, Ed. New York: John Wiley & Sons, Inc., Oct. 2001, pp. 123–173.

[120] E. A. Wan and R. van der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proceedings of the IEEE Adaptive Systems for Signal Processing, Communications, and Control Symposium*, 2000.

[121] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[122] O. Cappé, S. J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential Monte Carlo," *Proceedings of the IEEE*, vol. 95, no. 5, pp. 899–924, 2007.

[123] F. Daum, "Nonlinear filters: beyond the Kalman filter," *IEEE A&E Systems Magazine*, vol. 20, pp. 57–69, Aug. 2005.

[124] H. P. Saal, J. Ting, and S. Vijayakumar, "Multimodal nonlinear filtering using Gauss-Hermite quadrature," in *Proceedings of the 21st European Conference on Machine Learning*, 2011, pp. 81–96.

[125] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, vol. 3, no. 8, pp. 1445–1450, 1965.

[126] M. Klaas, M. Briers, N. de Freitas, A. Doucet, S. Maskell, and D. Lang, "Fast particle smoothing: if I had a million particles," in *Proceedings of the International Conference on Machine Learning*, 2006, pp. 25–29.

[127] S. Sarkka, "Unscented Rauch-Tung-Striebel smoother," *IEEE Transactions on Automatic Control*, vol. 53, no. 3, pp. 845–849, 2008.

[128] S. J. Godsill, A. Doucet, and M. West, "Monte Carlo smoothing for nonlinear time series," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 156–168, 2004.

[129] M. J. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, UK, 2003.

[130] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, vol. 3, no. 4, pp. 253–264, 1982.

[131] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions.* New York: Wiley, 1997.

[132] F. Dellaert, "The expectation maximization algorithm," Georgia Institute of Technology, Tech. Rep. GIT-GVU-02-20, 2002.

[133] C. M. Bishop, *Pattern Recognition and Machine Learning.* New York: Springer, 2006.

[134] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed., Cambridge, MA: MIT Press, 1998, pp. 355–368.

[135] C. F. J. Wu, "On the convergence properties of the EM algorithm," *The Annals of Statistics*, vol. 11, no. 1, pp. 95–103, 1983.

[136] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 1999, pp. 21–30.

[137] ——, "A variational Bayesian framework for graphical models," in *Advances in Neural Information Processing Systems*, vol. 12, 2000, pp. 209–215.

[138] M. J. Beal, F. Falciani, Z. Ghahramani, C. Rangel, and D. L. Wild, "A Bayesian approach to reconstructing genetic regulatory networks with hidden factors," *Bioinformatics*, vol. 21, no. 3, p. 349, 2005.

[139] G. Sanguinetti, N. D. Lawrence, and M. Rattray, "Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities," *Bioinformatics*, vol. 22, no. 22, pp. 2775–2781, 2006.

[140] N. L. J. Vermaak and P. Perez, "Variational inference for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 773–780.

[141] A. T. Cemgil, C. Févotte, and S. J. Godsill, "Variational and stochastic inference for Bayesian source separation," *Digital Signal Processing*, vol. 17, no. 5, pp. 891–913, 2007.

[142] W. Penny, S. Kiebel, and K. Friston, "Variational Bayesian inference for fMRI time series," *NeuroImage*, vol. 19, no. 3, pp. 727–741, 2003.

[143] V. Šmídl, "Variational Bayesian filtering," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 5020–5030, Oct. 2008.

[144] M. J. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," in *Bayesian Statistics 7*, J. M. Bernardo et al., Ed. Oxford: Clarendon Press, 2003, pp. 453–463.

[145] R. Milito, C. S. Padilla, R. A. Padilla, and D. Cadorin, "An innovations approach to dual control," *IEEE Transactions on Automatic Control*, vol. AC-27, no. 1, pp. 132–137, 1982.

[146] S. Fabri and V. Kadirkamanathan, *Functional Adaptive Control: an Intelligent Systems Approach.* London: Springer-Verlag, 2001.

[147] C. P. Robert and G. Casella, *Introducing Monte Carlo Methods with R.* New York: Springer-Verlag, 2010.

[148] M. A. Suchard, Q. Wang, C. Chan, J. Frelinger, A. Cron, and M. West, "Understanding GPU programming for statistical computation: studies in massively parallel massive mixtures," *Journal of Computational and Graphical Statistics*, vol. 19, no. 2, pp. 419–438, 2010.

[149] A. Lee, C. Yau, M. B. Giles, A. Doucet, and C. C. Holmes, "On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods," *Journal of Computational and Graphical Statistics*, vol. 19, no. 4, pp. 769–789, 2010.

[150] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[151] A. E. Gelfand, S. E. Hills, and A. Racine-Poon, "Illustration of Bayesian inference in normal data models using Gibbs sampling," *Journal of the American Statistical Association*, vol. 85, no. 412, pp. 972–985, 1990.

[152] C. K. Carter and R. Kohn, "On Gibbs sampling for state space models," *Biometrika*, vol. 81, no. 3, pp. 541–553, 1994.

[153] J. Geweke and H. Tanizaki, "Bayesian estimation of state-space models using the Metropolis-Hasting algorithm within Gibbs sampling," *Computational Statistics and Data Analysis*, vol. 37, no. 2, pp. 151–170, 2001.

[154] D. J. C. MacKay, "Introduction to Monte Carlo methods," in *Learning in Graphical Models*, M. I. Jordan, Ed. Cambridge, MA: MIT Press, 1998, pp. 175–204.

[155] K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, and W. Penny, "Variational free energy and the Laplace approximation," *NeuroImage*, vol. 34, pp. 220–234, 2007.

[156] T. J. L. K. Saul and M. I. Jordan, "Mean field theory for sigmoid belief networks," *Journal of Artificial Intelligence Research*, vol. 4, pp. 61–76, 1996.

[157] B. Wang and D. M. Titterington, "Inadequacy of interval estimates corresponding to variational Bayesian approximations," in *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, R. G. Cowell and Z. Ghahramani, Eds. Society for Artificial Intelligence and Statistics, 2005, pp. 373–380.

[158] T. G. Müller and J. Timmer, "Parameter identification techniques for partial differential equations," *International Journal of Bifurcation and Chaos*, vol. 14, no. 6, pp. 2053–2060, 2004.

[159] S. A. Sallberg, P. S. Maybeck, and M. E. Oxley, "Infinite-dimensional sampled-data Kalman filtering and the stochastic heat equation," in *Proceedings of the IEEE Conference on Decision and Control*, 2010, pp. 5062–5067.

[160] H. L. Choi, "Adaptive Sampling and Forecasting with Mobile Sensor Networks," Ph.D. dissertation, MIT, 2009.

[161] M. A. Demetriou and I. I. Hussein, "Estimation of spatially distributed processes using mobile spatially distributed sensor network," *SIAM Journal on Control and Optimisation*, vol. 48, no. 1, pp. 266–291, 2009.

[162] L. M. Berliner, C. K. Wikle, and N. Cressie, "Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling," *Journal of Climate*, vol. 13, no. 22, pp. 3953–3968, Nov. 2000.

[163] C. Wikle and N. Cressie, "A dimension-reduced approach to space-time Kalman filtering," *Biometrika*, vol. 86, no. 4, pp. 815–829, Dec. 1999.

[164] N. A. C. Cressie and C. K. Wikle, *Statistics for Spatio-temporal Data*. New Jersey: Wiley, 2011.

[165] G. Apaydin, S. Seker, and N. Ari, "Weighted extended b-splines for one-dimensional electromagnetic problems," *Applied mathematics and computation*, vol. 190, no. 2, pp. 1125–1135, 2007.

[166] M. A. Dewar, "A Framework for Dynamic Modelling of Spatiotemporal Systems," Ph.D. dissertation, University of Sheffield, 2006.

[167] J. Park and I. Sandberg, “Universal approximation using radial-basis-function networks,” *Neural Computation*, vol. 3, pp. 246–257, 1991.

[168] C. Guestrin, A. Krause, and A. P. Singh, “Near-optimal sensor placements in Gaussian processes,” in *Proceedings of the 22nd International Conference on Machine Learning*, 2005, pp. 265–272.

[169] A. J. Storkey, “Truncated covariance matrices and Toeplitz methods in Gaussian processes,” in *Proceedings of the International Conference on Artificial Neural Networks*, vol. 1, 1999, pp. 55–60.

[170] R. M. Sanner and J. J. E. Slotine, “Gaussian networks for direct adaptive control,” *IEEE Transactions on Neural Networks*, vol. 3, no. 6, pp. 837–863, 1992.

[171] T. A. Louis, “Finding the observed information matrix when using the EM algorithm,” *Journal of the Royal Statistical Society B*, vol. 44, pp. 226–233, 1982.

[172] X. L. Meng and D. B. Rubin, “Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm,” *Journal of the American Statistical Association*, vol. 86, no. 416, pp. 899–909, Dec. 1991.

[173] J. C. Duan and A. Fulop, “A stable estimator for the information matrix under EM for dependent data,” *Statistics and Computing*, vol. 21, no. 1, pp. 83–91, 2011.

[174] B. Efron and D. V. Hinkley, “Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information,” *Biometrika Trust*, vol. 65, no. 3, pp. 457–482, Dec. 1978.

[175] W. K. Newey and K. D. West, “A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix,” *Econometrica*, vol. 55, no. 3, pp. 703–708, 1987.

[176] M. J. Beal and Z. Ghahramani, “The variational Kalman smoother,” Gatsby Computational Neuroscience Unit, University College London, Tech. Rep. GCNU TR 2001-003, 2001.

[177] D. S. Watkins, *Fundamentals of Matrix Computations.* New York: Wiley-Interscience, 2002.

[178] J. R. Cannon, J. Douglas, and B. F. Jones, “Determination of the diffusivity of an isotropic medium,” *International Journal of Engineering Science*, vol. 1, no. 4, pp. 453–455, 1963.

[179] A. G. Madera, “Modelling of stochastic heat transfer in a solid,” *Applied Mathematical Modelling*, vol. 17, no. 12, pp. 664–668, 1993.

[180] B. Wahlberg and L. Ljung, “Design variables for bias distribution in transfer function estimation,” *IEEE Transactions on Automatic Control*, vol. 31, no. 2, pp. 134–144, 1986.

[181] I. Johnstone, “High dimensional bernstein-von mises: simple examples,” *Institute of Mathematical Statistics collections*, vol. 6, p. 87, 2010.

[182] A. Gelman, *Bayesian Data Analysis.* CRC press, 2004.

[183] T. J. R. Hughes, *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis.* New Jersey: Prentice Hall, 1987.

[184] J. Daunizeau, K. J. Friston, and S. J. Kiebel, “Variational bayesian identification and prediction of stochastic nonlinear dynamic causal models,” *Physica D: Nonlinear Phenomena*, vol. 238, no. 21, pp. 2089–2118, 2009.

[185] J. Møller, A. R. Syversveen, and R. P. Waagepetersen, “Log Gaussian Cox processes,” *Scandinavian Journal of Statistics*, vol. 25, no. 3, pp. 451–482, 1998.

[186] K. Yuan and M. Niranjan, “Estimating a state-space model from point process observations: a note on convergence,” *Neural Computation*, vol. 22, no. 8, pp. 1993–2001, Aug. 2010.

[187] E. N. Brown, R. Barbieri, U. T. Eden, and L. M. Frank, “Likelihood methods for neural spike train data analysis,” in *Computational Neuroscience: a Comprehensive Approach*, J. Feng, Ed. London: CRC Press, 2003, pp. 253–286.

[188] A. C. Smith *et al.*, “A Bayesian statistical analysis of behavioral facilitation associated with deep brain stimulation,” *Journal of Neuroscience Methods*, vol. 183, no. 2, pp. 267–276, 2009.

[189] M. Opper and C. Archambeau, “The variational Gaussian approximation revisited,” *Neural Computation*, vol. 21, no. 3, pp. 786–792, 2009.

[190] S. Koyama, L. Castellanos Pérez-Bolde, C. R. Shalizi, and R. E. Kass, “Approximate methods for state-space models,” *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 170–180, 2010.

[191] E. N. Brown, R. Barbieri, V. Ventura, R. E. Kass, and L. M. Frank, “The time-

rescaling theorem and its application to neural spike train data analysis," *Neural Computation*, vol. 14, no. 2, pp. 325–346, Feb. 2002.

[192] A. Riehle, S. Grün, M. Diesmann, and A. Aertsen, "Spike synchronization and rate modulation differentially involved in motor cortical function," *Science*, vol. 278, no. 5345, pp. 1950–1953, 1997.

[193] J. Stevens, *Applied Multivariate Statistics for the Social Sciences*, 4th ed. New Jersey: Lawrence Erlbaum Associates, 2002.

[194] D. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Process*, 2nd ed. New York: Springer-Verlag, 2003.

[195] R. N. Bracewell, *The Fourier Transform & its Applications*, 3rd ed. Singapore: McGraw-Hill, 2000.

[196] T. Minka, "Old and new matrix algebra useful for statistics [online]," http://research.microsoft.com/~minka/papers/matrix/, [Online; accessed 28-October-2010].

[197] L. Aniţa and S. Aniţa, "Note on the stabilization of a reaction-diffusion model in epidemiology," *Nonlinear Analysis: Real World Applications*, vol. 6, no. 3, pp. 537–544, Jul. 2005.

[198] P. A. Lewis and G. S. Shedler, "Simulation of nonhomogeneous Poisson processes by thinning," *Naval Research Logistics Quarterly*, vol. 26, no. 3, pp. 403–413, 1979.

[199] C. M. Bishop and I. T. Nabney, "NETLAB online reference documentation," https://ccrma.stanford.edu/~unjung/nethelp/index.htm, [Online; accessed 14-November-2011].

[200] P. J. Diggle, I. Kaimi, and R. Abellana, "Partial-likelihood analysis of spatio-temporal point-process data," *Biometrics*, vol. 66, pp. 347–354, 2010.

[201] O. Cappé and E. Moulines, "On-line expectation-maximization algorithm for latent data models," *Journal of the Royal Statistical Society Series B*, vol. 71, no. 3, pp. 593–613, Jun. 2009.

[202] V. Krishnamurthy and J. B. Moore, "On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure," *IEEE Transactions on Signal Processing*, vol. 41, no. 8, pp. 2557–2573, Aug. 1993.

[203] E. Weinstein, M. Feder, and A. V. Oppenheim, “Sequential algorithms for parameter estimation based on the Kullback-Leibler information measure,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 38, no. 9, pp. 1652–1654, 1990.

[204] H. Sohn *et al.*, “A review of structural health monitoring literature: 1996-2001,” Los Alamos National Laboratory, Los Alamos, NM, Tech. Rep. LA-13976-MS, 2004.

[205] E. A. Wan, R. van der Merwe, and A. T. Nelson, “Dual estimation and the unscented transformation,” in *Advances in Neural Information Processing Systems 12*, 2000, pp. 666–672.

[206] E. Wan and A. Nelson, “Dual Kalman filtering methods for nonlinear prediction, estimation, and smoothing,” in *Advances in Neural Information Processing Systems 9*, 1997.

[207] H. Moradkhani, S. Sorooshian, H. V. Gupta, and P. R. Houser, “Dual state-parameter estimation of hydrological models using ensemble Kalman filter,” *Advances in Water Resources*, vol. 28, no. 2, pp. 135–147, 2005.

[208] L. Nelson and E. Stear, “The simultaneous on-line estimation of parameters and states in linear systems,” *IEEE Transactions on Automatic Control*, vol. 21, no. 1, pp. 94–98, 1976.

[209] S. B. Chitralekha, J. Prakash, H. Raghavan, R. B. Gopaluni, and S. L. Shah, “A comparison of simultaneous state and parameter estimation schemes for a continuous fermentor reactor,” *Journal of Process Control*, vol. 20, no. 8, pp. 934–943, 2010.

[210] G. Kitagawa, “A self-organizing state-space model,” *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 1203–1215, Sep. 1998.

[211] G. Storvik, “Particle filters for state-space models with the presence of unknown static parameters,” *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 281–289, Feb. 2002.

[212] M. Bocquet, C. A. Pires, and L. Wu, “Beyond Gaussian statistical modeling in geophysical data assimilation,” *Monthly Weather Review*, vol. 138, no. 8, pp. 2997–3023, Aug. 2010.

[213] C. Andrieu and A. Doucet, “Online expectation-maximization type algorithms for parameter estimation in general state space models,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6, 2003, pp. 69–72.

[214] O. Cappé, “Online EM algorithm for hidden Markov models,” http://arxiv.org/abs/0908.2359, 2009, [Online; accessed 14-November-2011].

[215] M. Sato, “On-line model selection based on the variational Bayes,” *Neural Computation*, vol. 13, no. 7, pp. 1649–1681, Jul. 2001.

[216] V. Šmídl and A. Quinn, “The variational Bayes approximation in Bayesian filtering,” in *Proceedings of the 31$^{st}$ IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006, pp. 2588–2591.

[217] S. Sarkka and A. Nummenmaa, “Recursive noise adaptive Kalman filtering by variational Bayesian approximations,” *IEEE Transactions on Automatic Control*, vol. 54, no. 3, pp. 596–600, 2009.

[218] V. Šmídl and A. Quinn, “The restricted variational Bayes approximation in Bayesian filtering,” in *Proceedings of the IEEE Nonlinear Statistical Signal Processing Workshop*, 2006, pp. 224–227.

[219] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[220] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, “Novel approach to nonlinear/non-Gaussian Bayesian state estimation,” in *IEE Proceedings on Radar and Signal Processing*, 1993.

[221] P. M. di Lorenzo and J. D. Victor, “Taste response variability and temporal coding in the nucleus of the solitary tract of the rat,” *Journal of Neurophysiology*, vol. 90, no. 3, pp. 1418–1431, 2003.

[222] A. T. Roussin, J. D. Victor, J. Chen, and P. M. D. Lorenzo, “Variability in responses and temporal coding of tastants of similar quality in the nucleus of the solitary tract of the rat,” *Journal of Neurophysiology*, vol. 99, no. 2, pp. 644–655, 2008.

[223] W. Fong and S. Godsill, “Sequential Monte Carlo simulation of dynamical models with slowly varying parameters: application to audio,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2002, pp. 1605–1608.

[224] H. A. P. Blom and Y. Bar-Shalom, "The interacting multiple model algorithm for systems with Markovian switching coefficients," *IEEE Transactions on Automatic Control*, vol. 33, no. 8, pp. 780–783, 1988.

[225] J. O'Loughlin, F. D. W. Witmer, A. M. Linke, and N. Thorwardson, "Peering into the fog of war: the geography of the Wikileaks Afghanistan war logs, 2004-2009," *Eurasian Geography and Economics*, vol. 51, no. 4, pp. 472–495, 2010.

[226] D. Rohde and D. E. Sanger, "How a 'good war' in Afghanistan went bad," http://www.nytimes.com/2007/08/12/world/asia/12afghan.html, 2007, [Online; accessed 11-June-2011].

[227] J. Bohannon, "Counting the dead in Afghanistan," *Science*, vol. 331, no. 6022, pp. 1256–1260, Mar. 2011.

[228] "ANSO quarterly data report Q.4 2010," http://www.ngosafety.org/2010crs.html, 2010, [Online; accessed 14-November-2011].

[229] M. F. (Major General), "State of the insurgency: trends, intentions and objectives. Unclassified report, International Security Assistance Force, Afghanistan," http://www.humansecuritygateway.com/documents/ISAF_StateOfTheInsurgency_22December09.pdf, 2009, [Online; accessed 14-November-2011].

[230] S. Weinberger, "Social science: web of war." *Nature*, vol. 471, no. 7340, p. 566, 2011.

[231] J. O'Loughlin, F. D. W. Witmer, and A. M. Linke, "The Afghanistan-Pakistan wars, 2008–2009: micro-geographies, conflict diffusion, and clusters of violence," *Eurasian Geography and Economics*, vol. 51, no. 4, pp. 437–471, 2010.

[232] Al Jazeera, "Kabul urges polls attacks blackout," http://english.aljazeera.net/news/asia/2009/08/200981821718308671.html, 2009, [Online; accessed 14-November-2011].

[233] D. Conway, "Wikileaks analysis," https://github.com/drewconway/WikiLeaks_Analysis/tree/master/images, [Online; accessed 14-November-2011].

[234] P. Diggle, "A kernel method for smoothing point process data," *Applied Statistics*, vol. 34, pp. 138–147, 1985.

[235] M. Dewar, "Visualisation of activity in Afghanistan using the Wikileaks data," http://vimeo.com/14200191, [Online; accessed 14-November-2011].

[236] M. McCormick, P. Allen, and A. Dant, “Afghanistan war logs: IED attacks on civilians, coalition and Afghan troops,” http://www.guardian.co.uk/world/datablog/interactive/2010/jul/26/ied-afghanistan-war-logs, [Online; accessed 14-November-2011].

[237] W. J. Reed and B. D. Hughes, “From gene families and genera to incomes and internet file sizes: why power laws are so common in nature,” *Physical Review E*, vol. 66, no. 6, p. 67103, 2002.

[238] J. Meyerle, M. Katt, and J. Gavrilis, “ Counterinsurgency on the ground in Afghanistan,” CNA’s Center for Strategic Studies, Tech. Rep., 2010.

[239] T. Noetzel, “Germany’s small war in Afghanistan: military learning amid politico-strategic inertia,” *Contemporary Security Policy*, vol. 31, no. 3, pp. 486–508, 2010.

[240] “ANSO quarterly data report Q.4 2009,” http://www.ngosafety.org/2009crs.html, 2009, [Online; accessed 14-November-2011].

[241] Y. Lou and P. D. Christofides, “Nonlinear feedback control of surface roughness using a stochastic PDE: design and application to a sputtering process,” *Industrial & Engineering Chemical Research*, vol. 45, pp. 7177–7189, 2006.

[242] T. Baldacchino, S. Anderson, and V. Kadirkamanathan, “Structure detection and parameter estimation for NARX models in a unified EM framework,” *Automatica*, 2011, in press.

[243] H. Rue and S. Martino, “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations,” *Journal of the Royal Statistical Society B*, vol. 71, pp. 319–392, 2009.

[244] H. Rue and H. Tjelmeland, “Fitting Gaussian Markov random fields to Gaussian fields,” *Scandinavian Journal of Statistics*, vol. 29, no. 1, pp. 31–49, 2002.

[245] I. I. Hussein, “Kalman filtering with optimal sensor motion planning,” in *Proceedings of the American Control Conference*, 2008, pp. 3548–3553.

[246] Computer Vision Central$^{\text{TM}}$, “Google awards research grant in mobile crop surveillance,” http://computervisioncentral.com/content/google-awards-research-grant-mobile-crop-surveillance, [Online; accessed 09-September-2011].

[247] D. Stoyan and H. Stoyan, *Fractals, Random shapes, and Point Fields: Methods of Geometrical Statistics*. New York: Wiley, 1994.

[248] W. Steeb, *Matrix Calculus and Kronecker Product with Applications and C++ Programs.* Singapore: World Scientific, 1997.