

Chapter 2

Theory

This chapter introduces and details the mathematical and statistical framework underpinning our conflict modeling strategy. It is intended to be self-contained as much as possible, but it does rest on some mathematical prerequisites, chiefly in probability theory and real analysis. There is a large literature in mathematics, statistics and engineering covering these mathematical foundations in great detail; we highlight in particular the books by da Prato and Zabczyk (1993); Anderson et al. (1979); Jazwinski (1970); Ross (2006), but this list is by no means comprehensive.

Our modeling approach exploits the Bayesian paradigm, which combines prior knowledge with observations to quantify uncertainty in predictions. We therefore start by discussing the *observation model* (likelihood) we employ. As discussed in the introduction, the data type we are primarily concerned with consists of event logs, spatial and temporal coordinates of conflict events. Point processes are a convenient mathematical framework to describe event data, and Sect. 2.1 provides a simple and self-contained introduction to the main features of this class of stochastic processes. Of particular importance is the concept of the *Cox process* since it enables us to incorporate prior knowledge through a secondary stochastic process governing the *intensity* of the conflict.

Section 2.2 then describes the class of stochastic dynamical processes we employ to model conflict intensity. We focus in particular on stochastic partial differential equations and stochastic integro-difference equations, explaining how these two approaches are related to each other and how they provide a flexible framework to describe complex spatio-temporal behaviors. Our aim is not to provide a comprehensive introduction to this rich field of mathematics, but to explain in an operational way how these models can be used in the context of conflict modeling. With this in mind, we describe in some detail finite-dimensional reductions of SPDEs/SIDEs, leading to an algorithmically convenient state-space form. Following this, Sect. 2.3 outlines the field and parameter estimation components of the state-space framework. In particular, we employ standard message-passing and iterative algorithms for implementing quick and efficient recursive routines.

Finally, Sect. 2.4 gives the algorithmic details for (i) dimensionality reduction and (ii) the (approximate) Bayesian inference procedure we employ to combine prior and likelihood to obtain posterior estimates. The former includes a review of standard non-parametric methods which are often extremely useful to guide the modeling process. The latter is important and non-trivial, as large-scale Bayesian inference in spatio-temporal point processes is a challenging computational problem. We describe in detail the variational approach introduced in Zammit-Mangion et al. (2012b) and used in Zammit-Mangion et al. (2012a). However, we stress that this is an open research field in computational statistics and machine learning, and other approaches to large-scale approximate inference in spatio-temporal processes do exist (Rue et al. 2009; Cseke and Heskes 2011; Cseke et al. 2013); to our knowledge, a comparative analysis of these algorithms in terms of accuracy and efficiency has not yet been carried out.

Spatio-temporal modeling is a broad and important field in applied science: as such, mathematical ideas closely related to the ones described here have underpinned a very considerable body of research, both methodological and applied. While this research contains many insights that could be valuable in conflict modeling, we cannot review the field of spatio-temporal modeling here, but refer the interested reader to the excellent recent book of Cressie and Wikle (2011) for a comprehensive reference list and introduction to the field.

2.1 Point Processes

Extracting information from a pattern of points is often essential to reason about the underlying *cause* of the observations. In 1854, a physician by the name of John Snow linked the source of a cholera outbreak to a public water pump by examining the pattern of cases on a map. Today point patterns constitute core data sets in most scientific disciplines. In neuroscience, neural spike trains are frequently characterized by the firing time, whilst in ecology data typically correspond to sightings of an invading or endangered species. Virtually all natural hazards such as earthquakes or floods can be (and frequently are) summarized as points. Whatever the application, the core questions asked are ‘What can we tell about the pattern?’ And, more importantly, ‘Based on what we have observed so far, what can we *infer* from the pattern?’.

Two related approaches for analyzing point patterns emerge from asking these questions: the first considers characteristics of the pattern such as randomness or regularity and clustering effects. Theory for such analysis is firmly rooted in non-parametric techniques and methods here are routinely used for visual and exploratory analysis. Inference from the data, however, requires a statistical description of the mechanism by which the points are generated. This leads us to point-process modeling, a general framework which can be used for inference from temporal, spatial and spatio-temporal point patterns. As will be seen, these two approaches complement each other and are by no means mutually exclusive to the analyst. In this section, we review some basic mathematical tools which are useful to formalize concepts such

as randomness in point patterns and intensity of a point process. This also forms the basis for the hierarchical modeling framework we will adopt throughout the book, by specifying the observation model for conflict events. We focus here on the main concepts; relevant proofs and rigorous definitions can be found in the references.

2.1.1 Random Point Patterns and the Poisson Process

The approach taken in this book is rooted in Bayesian statistical modeling; Bayesian modeling quantifies uncertainty in predictions by treating all relevant variables as random variables and modeling the dependencies as conditional distributions. But how can we quantify randomness in a point pattern? We closely model our reasoning on the classic book by Kingman (1992). Let us consider the simple situation where the domain of interest (the space where the points appear) is the interval $[0, 1] \in \mathbb{R}$. In the case where we have a single point, the natural notion of randomness is the uniform distribution, i.e. every location within the interval has equal probability of being the point's coordinate. Formally, we may say that the *probability of finding the point in a subset A is proportional to the length (measure) of the subset*.

To generalize this notion of randomness to multiple points, we define a *generative process*: we assume that the position of each point is uniformly distributed within the interval and independent of all the other points. What is then the probability of finding k points in a subset A of measure $\mu(A)$? If one divides A into small bins (such that no bin contains more than one point), then since the position of the points are uniform in A , each bin has the same probability $\mu(A)/n$ of containing one point. Let the number of observed points be Y . Then the probability of observing k points, under this discretization, is

$$P(Y = k) = \binom{n}{k} \left(\frac{\mu(A)}{n} \right)^k \left(1 - \frac{\mu(A)}{n} \right)^{n-k} \quad (2.1)$$

where the leftmost combination term arises because the points are indistinguishable. As the bins are made smaller (and tend to zero in size) the number of bins $n \rightarrow \infty$. In this limit it can be shown, using standard algebra, that the probability of finding k points in a subset of measure $\mu(A)$ follows a *Poisson distribution with mean $\mu(A)$* ; i.e. $Pr(Y = k) = (\mu(A))^k e^{-\mu(A)} / k!$.

The generative process we have just described is the *Poisson point process*: the underlying uniformity is sometimes referred to as *spatial whiteness* or *complete randomness*. However, in many cases point patterns do not exhibit complete randomness: on the contrary, much of the valuable information which can be extracted from such data sets relies on the presence of well defined patterns such as clusters. These more complex patterns can to some extent be accommodated with a minor tweak of the generative process: we assume that each point is no longer drawn uniformly within the interval, but it comes from some (fixed) probability measure over the interval.

This is equivalent to *locally rescaling* the measure on the space by the so called *intensity* function, a non-negative function which measures the propensity for a point to happen in the neighborhood of a certain location.¹ Formally, the intensity function can be defined through a limiting procedure: let s denote a position in a measurable space V with measure μ , and let B denote a neighborhood of s . The intensity function $\lambda: V \rightarrow \mathbb{R}^+$ is given by

$$\lambda(s) = \lim_{\mu(B) \rightarrow 0} \frac{\Pr(\text{event in } B)}{\mu(B)}, \quad (2.2)$$

where \mathbb{R}^+ denotes the non-negative real numbers. Equation (2.2) is possibly the simplest model describing a given spatial point pattern. It can be easily shown that specification (2.2) implies that the number of events in a set A , $N(A)$, is *Poisson* distributed with a mean which is equal to the integrated intensity in the interval, i.e.

$$N(A) \sim \text{Poiss} \left(\int_A \lambda(s) ds \right). \quad (2.3)$$

This integrated intensity, also known as the *rate*, is only interval dependent: the number of events in two disjoint sets $N(A)$ and $N(B)$ are independent random variables. Note that for a constant intensity function $\lambda(s) = \lambda$, the expected number of points is again proportional to the area (or volume) of the region.

2.1.2 The Cox Process

Estimating the intensity $\lambda(s)$ from the observed events is one of the main tasks in point-process modeling. In many cases, however, a deterministic specification for $\lambda(s)$ is not sufficiently flexible for modeling data in practice. For example, we may want the numbers of points in disjoint sets to be correlated, or we may want the intensity function to depend noisily on some auxiliary information (covariates): for a conflict modeling example, we may wish to enforce that the conflict intensity is influenced by population density, or terrain type, both of which can be available information. Furthermore, from a conceptual point of view, our approach relies on quantifying uncertainty at all stages of the modeling process, including the estimation of the intensity function.

This is remedied by treating the intensity itself as the realization of a random field. In this setup one is then able to infer statistical properties of $\lambda(s)$ conditioned on the observed point pattern \mathcal{Y} . These *doubly stochastic* processes (randomness in $\lambda(s)$ and \mathcal{Y}), or *Cox* processes, constitute a powerful tool in point-process modeling. We are therefore naturally led to consider probability distributions

¹ We only consider the slightly restrictive assumption that the measure admits a density.

over spaces of functions; here, the main player is the *Gaussian Process*, a natural infinite-dimensional generalization of the Gaussian distribution.

Definition 2.1 (*Gaussian process* (e.g. Rasmussen and Williams 2006, Sect. 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 $\mu(s)$ and its covariance function $\Sigma(s, r)$ which for a real function $f(s)$ are given as $\mu(s) = \mathbb{E}[f(s)]$ and $\Sigma(s, r) = \mathbb{E}[(f(s) - \mu(s))(f(r) - \mu(r))]$. A draw $\varepsilon(s)$ from the GP is denoted as

$$\varepsilon(s) \sim \mathcal{GP}(\mu(s), \Sigma(s, r)). \quad (2.4)$$

An important special case of a Cox process is when the log-intensity function $z(s) = \ln(\lambda(s))$ is distributed according to a Gaussian Process.² In this setting the process is known as a log-Gaussian Cox process (LGCP). LGCPs and associated inference problems have been extensively studied (e.g. Møller and Waagepetersen 2004) in a number of application domains. From the modeling point of view, their main attraction consists in the possibility of explicitly incorporating knowledge about the intensity function (e.g. in the form of a specification of its dynamical behavior); Sect. 2.2 describes two ways in which such knowledge can be encoded. Before turning to the problem of modeling the intensity function, however, it is important to establish how a Poisson process can be practically used as a likelihood model.

2.1.3 The Poisson Process Likelihood Function

We have seen that Poisson and Cox processes provide a natural generative model for point data. While both these processes are non-parametric, infinite-dimensional mathematical objects, it turns out that they provide a surprisingly simple way of associating a likelihood to a set of observed points, turning them into a practical inference tool. Let \mathcal{Y} be a set of points from a (conditionally) Poisson process on $\mathcal{O} \subset \mathbb{R}^2$ with intensity function $\lambda(s)$. The Poisson process likelihood is given by (Kingman (1992); Møller and Waagepetersen (2004))

$$p(\mathcal{Y}|\lambda(s)) = \left(\prod_{s_j \in \mathcal{Y}} \lambda(s_j) \right) \exp \left(- \int_{\mathcal{O}} \lambda(s) ds \right). \quad (2.5)$$

To gain some understanding of the likelihood function (2.5), we split it into two components, $p(\mathcal{Y}|\lambda(s)) = L_1 L_2$ where $L_1 = \prod_{s_j \in \mathcal{Y}} \lambda(s_j)$ and $L_2 = \exp \left(- \int_{\mathcal{O}} \lambda(s) ds \right)$. The first component, L_1 is high when $\lambda(s)$ is large at observed event locations. The second component, L_2 on the other hand penalizes for over-all large $\lambda(s)$. The maximum likelihood solution is therefore one which promotes

² The logarithm is used to ensure positivity of the resulting intensity function.

a large $\lambda(s)$ only in regions where many events take place, subject to a smoothing regularizing penalty.

2.2 Spatio-Temporal Dynamic Models

This section briefly reviews two of the most common dynamic spatio-temporal models, the SPDE and the SIDE. In order to keep the exposition simple, in this book we will consider only linear spatio-temporal models. However both SPDEs and SIDEs are very general and flexible models which can accommodate non-linearities in a natural way. Nevertheless, linear spatio-temporal models can already produce complex behaviors, and, given the considerable statistical difficulties posed by non-linear models, linear models often give a good compromise between tractability and mechanistic detail.

2.2.1 Partial Differential Equations and Their Stochastic Counterpart

Partial differential equations (PDEs) are continuous-time continuous-space models which have been used extensively to describe a wide range of natural and engineered systems. The best known example of PDE is probably the *heat equation*, governing the spatio-temporal evolution of the temperature of a body. This is derived directly from Fourier's law via the divergence theorem, and takes the form

$$\frac{\partial T(s, t)}{\partial t} = \frac{D \partial^2 T(s, t)}{\partial s^2}, \quad s \in [a, b], \quad (2.6)$$

where $T(s, t)$ is the temperature field, $[a, b]$ is the region occupied by the body (assumed for simplicity to be one dimensional) and D is the heat diffusion parameter. In order to solve this equation, one must specify two sets of conditions: the *initial conditions* $T(s, t = 0)$, specifying the temperature field at the initial time of interest at every spatial location, and the *boundary conditions*. The boundary conditions specify the behavior of the field at the extremes of the domain: typically, one assumes that either the field must be constant at the boundaries (*Dirichlet conditions*), or that the derivative of the field must be constant at the boundaries (*Neumann conditions*; in the heat equation example, a vanishing derivative corresponds to a *thermally insulated* body).

Formally, a PDE is defined as any equation which involves an unknown function of two or more independent variables and one or more of its partial derivatives (Evans 1998, Sect. 1.1). In spatio-temporal systems the independent variables are restricted to be space and time respectively. Let $z(s, t)$ be a one-dimension spatio-temporal field for simplicity. Then the general form of the PDE is given by

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

If $F(\cdot)$ is a linear functional 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.

While PDEs are extremely flexible modeling tools, their deterministic behavior limits their usefulness in modeling random processes. To obviate this problem, a stochastic term needs to be added to the deterministic PDE (e.g. Dalang and Frangos 1998; Carmona 1998 Sect.1.1); the simplest example is the addition of a spatio-temporal Gaussian noise process. The resulting class of models, the SPDE, constitutes one of the most powerful and flexible tools for spatio-temporal modeling. Prévôt and Röckner (2007), Chap. 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.

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

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

where $z(s, t)$ is the random field, $D(s) > 0$, $\sigma \in \mathbb{R}^+$ and $\dot{W}(s, t)$ is space-time noise. An appealing feature of SPDEs is the ease with which spatial heterogeneity can be accommodated—for instance note how spatially heterogeneous dynamics are immediately apparent in the spatially varying parameter $D(s)$ which moreover retains physical meaning. This spatially varying diffusion can be useful in conflict modeling e.g. to model variable conflict dissipation along a linear segment (Zhukov 2012).

2.2.2 Integro-Difference Equation Models

Integro-difference equation (IDE) models are powerful constructs for describing spatio-temporal behavior. The deterministic flavor was introduced by Kot and Schaffer (1986) as a tool to model the spread of invading organisms. Let $s \subset \mathcal{O}$ and k denote continuous space and a discrete-time index respectively, $z_k(s)$ be the spatio-temporal field under study and \mathcal{O} the domain of interest. Then the IDE is given by

$$z_{k+1}(s) = \mathcal{A} f(z_k(s)) := \int_{\mathcal{O}} k_I(s, \mathbf{r}) f(z_k(\mathbf{r})) d\mathbf{r}, \quad (2.9)$$

where $k_I(\cdot, \cdot)$ is a mixing kernel and $f(\cdot)$ is a one-to-one mapping. As with the SPDE in Sect. 2.2.1, there are strong conceptual and practical reasons to prefer a stochastic

treatment of the intensity field. We will therefore consider the stochastic IDE (Wikle 2002) which incorporates additive spatial noise (in the form of a spatial GP) in the IDE. In the SIDE, at each time step the propagated field is superimposed with a draw from a spatial GP, $e_k(\mathbf{s}) \sim \mathcal{GP}(\mu_e(\mathbf{s}), k_e(\mathbf{s}, \mathbf{r}))$ where $\mu_e(\mathbf{s})$ is a spatially resolved trend and $k_e(\mathbf{s}, \mathbf{r})$ is the covariance function. In compact form the evolution equation of the SIDE is then given as

$$z_{k+1}(\mathbf{s}) = \mathcal{A}z_k(\mathbf{s}) + e_k(\mathbf{s}). \quad (2.10)$$

The (S)IDE may be split into two stages, the *sedentary stage* and the *dispersion stage*. Together with the additive disturbance, these two mechanisms wholly describe the evolution of the spatio-temporal field. The sedentary stage is exercised through the mapping $f(\cdot)$ which models the *local* behavior of the field in time. It seeks to model local growth or decay; for example in ecology the standard logistic or Ricker growth models are frequently used (Kot and Schaffer 1986). In this book, for simplicity, we take the linear case $f(z_k(\mathbf{r})) = \gamma z_k(\mathbf{r})$ with $\gamma = 1$ (Dewar et al. 2009). The second, dispersion, stage models the field dynamics and seeks to answer the question ‘How does the field dissipate or relocate (migrate) in time?’. The integral operator \mathcal{A} , together with the mixing kernel $k_I(\mathbf{s}, \mathbf{r})$ are used to encode this behavior. For simplicity we assume again that $f(\cdot)$ is linear, and let the kernel depend solely on $v = \|\mathbf{s} - \mathbf{r}\|$. Then the dynamics are spatially invariant and the field is termed *homogeneous* and *isotropic* (rotationally invariant). In this case, \mathcal{A} reduces to the well-known *convolution* operator.

The kernel of a convolution operator determines many visually immediate properties of the generated fields. For example, if $k_I(v)$ has a negative lobe at some distance v^* , then the spatio-temporal interaction for separations v^* is inhibitory. On the other hand if the center of mass is not at the origin, then mobility/advection is modelled. Indeed, the greatest attraction of the IDE is that the kernel can give substantial insight into system behavior, making the IDE a very flexible and intuitive tool for spatio-temporal modeling. Figure 2.1 shows, for instance, how a kernel with an offset center of mass produces random fields with a clear sense of directionality. Obviously the kernel is rarely known and statistically the main challenge is to estimate $k_I(\mathbf{s}, \mathbf{r})$. This was the focus of several works in recent years (Dewar et al. 2009; Zammit Mangion et al. 2011a; Freestone et al. 2011).

The SIDE has been successfully used in a vast range of applications, from cloud intensity modeling (Wikle 2002) to electroencephalography signal modeling (Freestone et al. 2011). In general, it provides a very flexible modeling framework for complex spatio-temporal systems.

2.2.2.1 Relationship Between SPDEs and SIDs

SPDEs and SIDs are clearly closely related classes of models: SPDEs focus on an infinitesimal, mechanistic description of the system, while SIDs provide an integrated, global version of the dynamics. This relationship is best seen by considering the example of the one-dimensional homogeneous heat equation with constant diffusion:

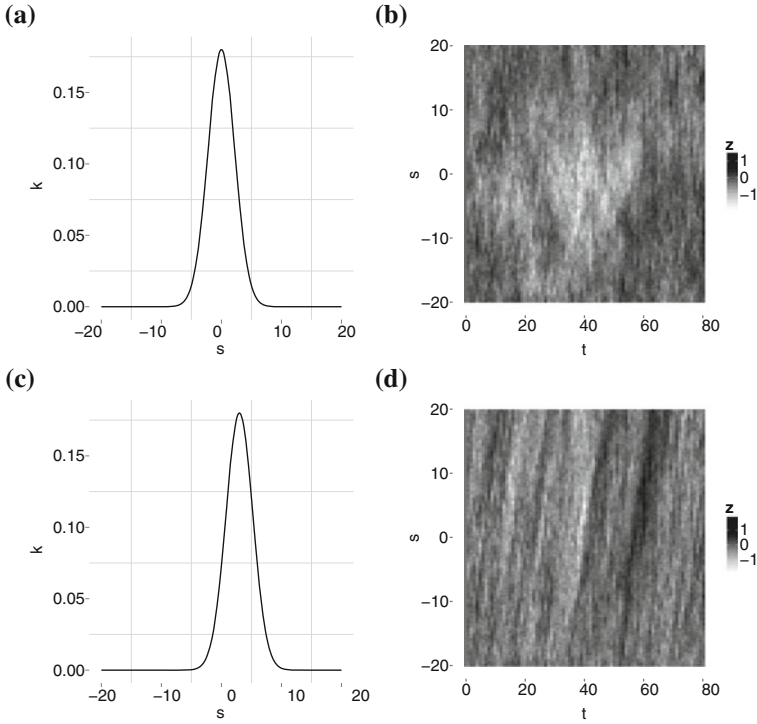


Fig. 2.1 Reflecting spatio-temporal patterns with the SIDE. **a** Centered mixing kernel. **b** Realization of SIDE field with shifted kernel. **c** Shifted kernel. **d** Realization of SIDE with shifted kernel

$$\frac{\partial z(s, t)}{\partial t} = \frac{D \partial^2 z(s, t)}{\partial s^2}, z(s, 0) = z_0(s). \quad (2.11)$$

It can be shown using Fourier analysis methods (Coleman 2005) that the solution of this PDE is

$$z(s, t) = \frac{1}{\sqrt{4\pi Dt}} \int_{\mathcal{O}} e^{\frac{-(s-s')^2}{4Dt}} z_0(s') ds'. \quad (2.12)$$

Thus the solution of the homogeneous heat equation is an IDE with a squared exponential function as the mixing kernel, $f(\cdot)$ a linear function and with initial condition $z_0(s)$. However, the constant D is now embedded within $f(\cdot)$ and $k_I(s, s')$ and the physical interpretation of relocation or diffusion is lost in the process. To recover this term it would be required to compare the IDE with the original PDE.

Another difference is the way spatial heterogeneity can be accommodated. As we have seen in Eq. (2.8), spatial heterogeneity can easily be handled in the SPDE by having spatially varying constants. This contrasts with the IDE where the heterogeneity

is implemented in the redistribution kernel based on observed spatio-temporal behavior (Wikle 2002). Both the SIDE and SPDE thus have their own modeling advantages and disadvantages. The choice between the two as the class of models of choice is not always a straightforward decision and is highly dependent on prior knowledge available from the system under study.

2.2.3 Dimensionality Reduction in Spatio-Temporal Models

Both the SIDE and the SPDE are continuous-space models, and hence infinite-dimensional constructs. For computational tractability it is of great benefit to approximate the models on a finite-dimensional space. There are a number of computational strategies that aim at reducing these systems into a form amenable to standard signal processing techniques, which are usually tailored for finite-dimensional systems. In this work we make use of the *method of moments*³ (Hausenblas 2003; Harrington 1993, Sect.1.3). Consider the simple linear equation $\mathcal{A}z(s) = f(s)$ for which it is required to find a solution for $z(s)$. By approximately expanding $z(s)$ as a series of n basis functions $\{\phi_i(s)\}_{i=1}^n$ with weights $x_1 \dots x_n$, one obtains an approximation $z(s) \approx \sum_{i=1}^n x_i \phi_i(s)$, see Fig. 2.2. Consequently

$$\sum_{i=1}^n x_i \mathcal{A} \phi_i(s) = f(s). \quad (2.13)$$

The method of moments proceeds by taking the inner product $\langle \cdot, \cdot \rangle$ of (2.13) with respect to a set of m testing functions $\{\chi_i(s)\}_{i=1}^m$ to obtain the set of equations

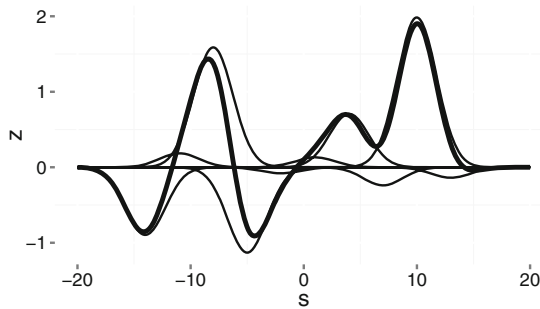


Fig. 2.2 Basis expansion of a function (*thick line*) as a linear combination of Gaussian radial basis functions (*thin lines*)

³ This is not to be confused with the method of moments associated with parameter estimation.

$$\sum_{i=1}^n x_i \langle \chi_j, \mathcal{A} \phi_i \rangle = \langle \chi_j, f \rangle, j = 1 \dots m. \quad (2.14)$$

Notice that this equation no longer depends on s , as the inner products involve integration over s . The set of equations may be written in matrix form to give

$$\mathbf{A} \mathbf{x} = \mathbf{f}, \quad (2.15)$$

where

$$\mathbf{A} = \begin{bmatrix} \langle \chi_1, \mathcal{A} \phi_1 \rangle & \langle \chi_1, \mathcal{A} \phi_2 \rangle & \dots & \langle \chi_1, \mathcal{A} \phi_n \rangle \\ \langle \chi_2, \mathcal{A} \phi_1 \rangle & \langle \chi_2, \mathcal{A} \phi_2 \rangle & \ddots & \langle \chi_2, \mathcal{A} \phi_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \chi_m, \mathcal{A} \phi_1 \rangle & \langle \chi_m, \mathcal{A} \phi_2 \rangle & \dots & \langle \chi_m, \mathcal{A} \phi_n \rangle \end{bmatrix} = [\langle \chi_i, \mathcal{A} \phi_j \rangle]_{i,j=1}^{m,n}, \quad (2.16)$$

and the vectors $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$, $\mathbf{f} = [\langle \chi_1, f \rangle, \langle \chi_2, f \rangle, \dots, \langle \chi_m, f \rangle]^T$. If \mathbf{A} is square and its inverse exists then the required solution is given by solving for \mathbf{x} using standard linear algebraic methods. 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, $\{\chi_i(s)\}_{i=1}^m = \{\phi_i(s)\}_{i=1}^n$, $m = n$. The Galerkin method has been extensively used for both SPDE (Hausenblas 2003) and SIDE (Dewar et al. 2009; Scerri 2010) approximation.

In the discrete-time SIDE case, the Galerkin method results in a linear, discrete-time *state-space model*. To see this, consider the standard SIDE of (2.10), expand $z_k(s)$ as

$$z_k(s) \approx \sum_{i=1}^n x_{k,i} \phi_i(s), \quad (2.17)$$

and further take the inner product with respect to $\{\phi_j(s)\}_{j=1}^n$ to obtain

$$\mathbf{x}_{k+1} = \Psi_{\mathbf{x}}^{-1} \Psi_{\mathcal{A}} \mathbf{x}_k + \Psi_{\mathbf{x}}^{-1} \langle \boldsymbol{\phi}(s), e_k(s) \rangle, \quad (2.18)$$

where

$$\Psi_{\mathbf{x}} = \langle \boldsymbol{\phi}(s), \boldsymbol{\phi}(s)^T \rangle, \quad (2.19)$$

$$\Psi_{\mathcal{A}} = \langle \boldsymbol{\phi}(s), \mathcal{A} \boldsymbol{\phi}(s)^T \rangle. \quad (2.20)$$

The matrix $\Psi_{\mathbf{x}}$ is the *Gram* matrix. The matrices $\Psi_{\mathbf{x}}$ and $\Psi_{\mathcal{A}}$ are related to the *mass* and *stiffness* matrices in finite-element decomposition.

With SPDEs, some form of temporal discretization is required in conjunction with the Galerkin method to obtain a discrete-time model of the form (2.18). In general this may be obtained through a *six-point* finite-difference scheme (Grossmann et al.

(2007, Sect. 2.6)) which may be defined through a user-defined parameter γ . For example, on a grid this scheme approximates the PDE (e.g. 2.6) to

$$\frac{z_{j,k+1} - z_{j,k}}{\Delta_t} = \mathcal{A}^N(\gamma z_{j,k+1} + (1 - \gamma)z_{j,k}), \quad j \in \mathbb{Z}, k \in \mathbb{Z}^+, \quad (2.21)$$

with initial condition $z_{j,0} = z_0(j\Delta_s)$ and where Δ_t is a fixed-width interval within the temporal domain and where \mathcal{A}^N is the finite-dimensional representation of the spatial differential operator \mathcal{A} . 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; see Hausenblas (2003) for properties relating to each of these schemes in an SPDE context.

Approximate noise process: With both SPDEs and SDEs, the statistical properties of the projected term $\mathbf{e}_k = \Psi_x^{-1} \langle \boldsymbol{\phi}(s), e_k(s) \rangle$ are required. Recall that we assume that the noise term $e_k(s)$ is drawn from a Gaussian process with mean $\mu_e(s)$ covariance function $k_e(s, r)$. By standard properties of GPs (e.g. Rasmussen and Williams 2006, Chap. 2), the projection of a sample from a GP on a finite set of basis functions is distributed according to a multivariate normal distribution. The statistics of this distribution are obtained through the standard rules of expectation

$$\mathbb{E}[\mathbf{e}_k] = \Psi_x^{-1} \langle \boldsymbol{\phi}(s), \mu_e(s) \rangle \quad (2.22)$$

$$\begin{aligned} \text{cov}[\mathbf{e}_k \mathbf{e}_k^T] &= \Psi_x^{-1} \text{cov}[\langle \boldsymbol{\phi}(s), e_k(s) \rangle \langle \boldsymbol{\phi}(r)^T, e_k(r) \rangle] \Psi_x^{-1} \\ &= \Psi_x^{-1} \left[\iint k_e(s, r) \boldsymbol{\phi}(s) \boldsymbol{\phi}(r)^T ds dr \right] \Psi_x^{-1}. \end{aligned} \quad (2.23)$$

The integrals in (2.22), (2.23) can be tedious to compute, especially when $\mu_e(s)$ and $k_e(s, r)$ are partially unknown in which case they require repeated evaluation in an estimation framework. Considerable simplifications may be achieved, however, by assuming that the mean and covariance function can themselves be decomposed as a sum of the spatial basis functions

$$\mu_e(s) \approx \boldsymbol{\phi}(s)^T \boldsymbol{\vartheta}, \quad (2.24)$$

$$k_e(s, r) \approx \boldsymbol{\phi}(s)^T \Sigma_e \boldsymbol{\phi}(r), \quad (2.25)$$

for some $\boldsymbol{\vartheta} \in \mathbb{R}^n$, $\Sigma_e \in \mathbb{R}^{n \times n}$. The mean and covariance of \mathbf{e}_k then simply reduce to $\boldsymbol{\vartheta}$ and Σ_e , which may be estimated (partially or wholly) easily within a standard estimation framework. This will be exploited in Chap. 3.

Approximate mixing kernel in the SIDE: Since the inner product in (2.20) may be hard to compute, it is beneficial to also find a finite-dimensional approximation of \mathcal{A} . We thus decompose $k_I(s, r)$ using a (usually much smaller) set of basis functions $\boldsymbol{\phi}_{k_I} \in \mathbb{R}^{n_{k_I}}$ to obtain

$$k_I(s, r) = \boldsymbol{\phi}_{k_I}(s)^T \Sigma_I \boldsymbol{\phi}_{k_I}(r), \quad (2.26)$$

where $\Sigma_I \in \mathbb{R}^{n_{k_I} \times n_{k_I}}$. Note that, as with the field decomposition, ϕ_{k_I} need not be orthonormal. This decomposition leads to a similar formulation of Dewar et al. (2009). Under this decomposition, $\Psi_{\mathcal{A}}$ in (2.20) is given by

$$\Psi_{\mathcal{A}} = \langle \phi(s), \mathcal{A} \phi(s)^T \rangle \quad (2.27)$$

$$= \left\langle \phi(s), \int_{\mathcal{O}} \phi_{k_I}(s)^T \Sigma_I \phi_{k_I}(r) \phi(r)^T dr \right\rangle \quad (2.28)$$

$$= \iint_{\mathcal{O}} \phi(s) \phi_{k_I}(s)^T \Sigma_I \phi_{k_I}(r) \phi(r)^T dr ds \quad (2.29)$$

$$= \langle \phi(s), \phi_{k_I}(s)^T \rangle \Sigma_I \langle \phi_{k_I}(s), \phi(s)^T \rangle, \quad (2.30)$$

to give

$$\Psi_{\mathcal{A}} = \Phi_{k_I} \Sigma_I \Phi_{k_I}^T, \quad (2.31)$$

where $\Phi_{k_I} = \langle \phi(s), \phi_{k_I}(s)^T \rangle$. If we let $\phi_{k_I}(s) = \phi(s)$ then $\Psi_{\mathcal{A}} = \Psi_x \Sigma_I \Psi_x$. Under these finite-dimensional approximations, the resulting representation of the SIDE is n dimensional and given by

$$\mathbf{x}_{k+1} = \Sigma_I \Psi_x \mathbf{x}_k + \mathbf{e}_k, \quad (2.32)$$

where $\mathbf{e}_k \sim \mathcal{N}(\boldsymbol{\vartheta}, \Sigma_e)$. In the special case of \mathcal{A} being the identity operator (i.e. a null dispersion stage), then no decomposition is required since $\Psi_{\mathcal{A}} = \Psi_x$. From (2.18) this yields a standard random-walk model which is convenient to work with. Note that spatio-temporal interactions are still present in this model if Σ_e has non-zero off-diagonal elements.

Approximate likelihood: So far we have discussed how the method of moments enables the reduction of the SPDE/ SIDE models to linear state-space models. In order to set up an inferential framework, we must now consider how the Galerkin projection affects the observation model in our framework, i.e. the conditionally Poisson point process. Applying the reduction in (2.17) to (2.5), i.e. at a single time frame, we obtain the following approximate likelihood function

$$p(\mathcal{Y}|\mathbf{x}) = \left(\prod_{s_j \in \mathcal{Y}} \exp(\phi^T(s_j)\mathbf{x}) \right) \exp \left(- \int_{\mathcal{O}} \exp(\phi^T(s)\mathbf{x}) ds \right) = L_1 \times L_2. \quad (2.33)$$

This again factorizes in two parts, one involving only the projection of the intensity function at the observed points, and the other involving the integral of the basis functions over the whole space. It is important to notice the implications of Eq. (2.33) in terms of inference. The first part of the likelihood, L_1 , is log-linear in \mathbf{x} and presents no difficulty in computing maximum likelihood or Bayes estimates. On the other hand, the second term L_2 contains an intractable double exponential, prohibiting an

analytical estimation of \mathbf{x} and motivating the use of approximation strategies (such as the ones detailed in the next section).

Considering now the spatio-temporal case, then, the single time-frame likelihood at k is

$$p(\mathcal{Y}_k|\mathbf{x}) = \left(\prod_{s_j \in \mathcal{Y}_k} \Delta_t \exp(\boldsymbol{\phi}^T(s_j)\mathbf{x}_k) \right) \exp \left(-\Delta_t \int_{\mathcal{O}} \exp(\boldsymbol{\phi}^T(s)\mathbf{x}_k) ds \right), \quad (2.34)$$

where Δ_t is the length of the observed time frame (usually 1 day or 1 week). Conditional on \mathbf{x}_k , each spatial point pattern is independent from any other in the time series. Thus, the spatio-temporal likelihood for a sequence of K equally spaced frames of width Δ_t is

$$p(\{\mathcal{Y}_k\}|\mathbf{x}_0 \dots \mathbf{x}_K) = \prod_{k=1}^K p(\mathcal{Y}_k|\mathbf{x}_k), \quad (2.35)$$

which will be the observation model within our hierarchical framework.

So far, we have presented the basic ideas behind dimensionality reduction in spatio-temporal systems. An important aspect which we have not discussed is the choice of the finite set of basis functions. Clearly, this is very important: an inappropriate choice may simply lead to interesting dynamics being projected out. To the best of our knowledge, there is no simple solution to this problem: the choice of basis function has to be considered carefully within the modeling process, and application-specific domain knowledge may be required. Section 2.4.2 describes in detail the non-parametric method in Zammit-Mangion et al. (2012b,a), which has the potential of being widely applicable as it relies on minimal assumptions.

2.3 Smoothing and Approximate Inference

Sections 2.1 and 2.2 introduced the likelihood and prior models we will consider in this book. In this section we instead focus on the inference problem of obtaining posterior estimates of the intensity function from point observations.

The scenario we are interested in can be characterized in the following way: our observations (conflict events) are an indirect, noisy result of an auxiliary, unobserved random variable (the conflict intensity). The crucial assumption, embedded in the Cox process formulation, is that the observations at different time points (and at different locations) are *independent* of each other *conditioned* on the intensity function, so that all the complex spatio-temporal dynamics displayed by the observations are entirely due to the dynamics of the intensity. Conditional independence relationships such as these are often depicted graphically in machine learning and

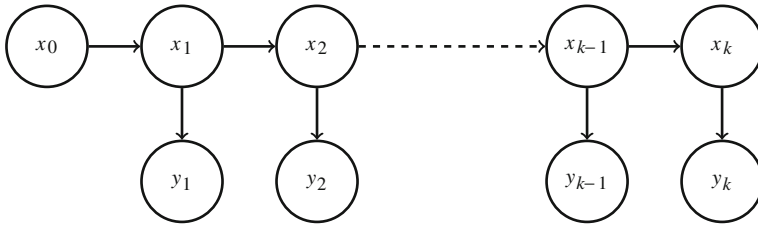


Fig. 2.3 Graphical representation of a state-space model showing the evolution of the latent states x_k and the observations y_k

statistics⁴; Fig. 2.3 shows the graphical model corresponding to the state-space models we consider in this book. This situation, where observations are conditionally independent given unobserved dynamical variables, is common in many engineering and scientific applications: for example, in speech recognition our observations (sound waveforms) are noisy *emissions* determined by underlying unobserved variables (the phonemes we intended to utter). Similar problems arise in domains ranging from bioinformatics to robotics and control. There exists therefore a large literature concerned with statistical inference in this class of models.

Traditionally, the underlying, unobserved variables are termed *states* in the engineering literature, hence the term state-space model. Conditioned on the states, the observations are independent. However, the dynamics of the states, and the distribution of the observations given the states, also depend on a number of parameters (for simplicity assumed to be static); these parameters also need to be estimated from data in general. The problems of estimating states and parameters are traditionally considered separately, even if they are in effect two sides of the same coin. We will start approaching the state estimation problem in Sect. 2.3.1. We will take a slightly more general approach, and describe in some detail an iterative algorithm to solve this problem for general state-space models, the *two-filter smoother* (Sect. 2.3.2). We then move on to consider the problem of jointly estimating states and parameters. In keeping with the general philosophy of this book, we attempt to quantify uncertainty both on states and parameters through the use of Bayes' theorem. An analytical solution for the posterior distribution is however not available for the class of models we consider; we therefore present a class of approximate inference algorithms based on a variational approach (Sect. 2.3.3).

⁴ The whole field of graphical models and graphical statistics is concerned with inference algorithms for probability distributions exhibiting specific conditional independence relationships, cf. Bishop (2006).

2.3.1 State Estimation

A discrete-time finite-dimensional state-space model consists of a real-valued state vector $\mathbf{x}_k \in \mathbb{R}^n$ following a first-order Markov process. The sequence of states is not observed directly, but rather, through observations $\mathbf{y}_k \in \mathbb{R}^m$, as depicted in Fig. 2.3.⁵ From the figure several conditional dependencies which facilitate algorithm derivation may be highlighted, for instance that $\mathbf{y}_k | \mathbf{x}_k$ is independent of $(\mathbf{x}_0 \dots \mathbf{x}_{k-1}, \mathbf{y}_1 \dots \mathbf{y}_{k-1})$ or that $\mathbf{x}_k | \mathbf{x}_{k-1}$ is independent of $(\mathbf{x}_0 \dots \mathbf{x}_{k-2}, \mathbf{y}_1 \dots \mathbf{y}_{k-1})$.

The general model under consideration in this book is

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{e}_k, \quad \mathbf{e}_k \sim \mathcal{N}_{\mathbf{e}_k}(\boldsymbol{\vartheta}, \Sigma_e), \quad (2.36)$$

$$p(\mathbf{y}_k | \mathbf{x}_k) = f(\mathbf{x}_k; \mathbf{b}), \quad (2.37)$$

where $\mathbf{A} = \Sigma_I \Psi_{\mathbf{x}} \in \mathbb{R}^{n \times n}$ is the state transition matrix and \mathbf{b} is a vector of parameters appearing in the observation model. Determining the states $\mathcal{X} = \mathbf{x}_{0:K} = \{\mathbf{x}_0 \dots \mathbf{x}_K\}$ from $\mathcal{Y} = \mathbf{y}_{1:K} = \{\mathbf{y}_1 \dots \mathbf{y}_K\}$ is known as the *state estimation* problem. The optimal estimation of \mathcal{X} from a data set is referred to as the *smoothing* problem, explored in Sect. 2.3.2. If, in addition to \mathcal{X} , parameters composing \mathbf{A} , Σ_w , $\boldsymbol{\vartheta}$, \mathbf{b} need to be estimated the problem is referred to as a *joint field-parameter estimation* problem. One method to solve this problem is variational Bayes expectation maximization (VBEM) discussed in Sect. 2.3.3.

2.3.2 Filtering and Smoothing

There are two widely accepted approaches for obtaining the posterior distribution of \mathbf{x}_k , i.e. the distribution of \mathbf{x}_k conditioned on the whole data set $p(\mathbf{x}_k | \mathcal{Y})$ (Briers et al. 2004). The first is the forward-backward algorithm in which a forward pass (also known as *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. Since the two-filter smoother is more amenable to variational techniques, here we restrict our discussion to the latter. Throughout the book, 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.

The two-filter smoother is a result of the factorization

⁵ For point processes we define \mathbf{y}_k as the spatial coordinates of points in \mathcal{X}_k .

$$\begin{aligned}
p(\mathbf{x}_k|\mathcal{Y}) &= p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{y}_{k+1:K}) \\
&= \frac{p(\mathbf{x}_k|\mathbf{y}_{1:k})p(\mathbf{y}_{k+1:K}|\mathbf{x}_k, \mathbf{y}_{1:k})}{p(\mathbf{y}_{k+1:K}|\mathbf{y}_{1:k})} \\
&\propto p(\mathbf{x}_k|\mathbf{y}_{1:k})p(\mathbf{y}_{k+1:K}|\mathbf{x}_k) \\
&= \alpha_k(\mathbf{x}_k)\beta_k(\mathbf{x}_k),
\end{aligned} \tag{2.38}$$

which is obtained by repeated application of Bayes' rule and by using the conditional independences implied by the graphical model in Fig. 2.3. $\alpha_k(\cdot)$ is called the *forward message* and $\beta_k(\cdot)$ is the *backward message*. The forward message is given as

$$\alpha_k(\mathbf{x}_k) = \frac{p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y}_{1:k-1})}{p(\mathbf{y}_k|\mathbf{y}_{1:k-1})} \propto p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y}_{1:k-1}), \tag{2.39}$$

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

$$p(\mathbf{x}_k|\mathbf{y}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{y}_{1:k-1})d\mathbf{x}_{k-1}. \tag{2.40}$$

Therefore

$$\begin{aligned}
\alpha_k(\mathbf{x}_k) &\propto p(\mathbf{y}_k|\mathbf{x}_k) \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{y}_{1:k-1})d\mathbf{x}_{k-1} \\
&= p(\mathbf{y}_k|\mathbf{x}_k) \int p(\mathbf{x}_k|\mathbf{x}_{k-1})\alpha_{k-1}(\mathbf{x}_{k-1})d\mathbf{x}_{k-1}.
\end{aligned} \tag{2.41}$$

The forward messages α_k can thus be found recursively by starting from an initial estimate $\alpha_0(\mathbf{x}_0)$ (it propagates forward information from the past towards the future). Similarly, the backward message is found from

$$\begin{aligned}
\beta_k(\mathbf{x}_k) &= \int p(\mathbf{y}_{k+1:K}, \mathbf{x}_{k+1}|\mathbf{x}_k)d\mathbf{x}_{k+1} \\
&= \int p(\mathbf{y}_{k+1:K}|\mathbf{x}_{k+1})p(\mathbf{x}_{k+1}|\mathbf{x}_k)d\mathbf{x}_{k+1} \\
&= \int p(\mathbf{y}_{k+1}|\mathbf{x}_{k+1})p(\mathbf{y}_{k+2:K}|\mathbf{x}_{k+1})p(\mathbf{x}_{k+1}|\mathbf{x}_k)d\mathbf{x}_{k+1} \\
&= \int p(\mathbf{y}_{k+1}|\mathbf{x}_{k+1})\beta_{k+1}(\mathbf{x}_{k+1})p(\mathbf{x}_{k+1}|\mathbf{x}_k)d\mathbf{x}_{k+1},
\end{aligned} \tag{2.42}$$

which can thus also be found recursively (hence the two-filter smoother is a recursive algorithm for state estimation). The marginal posterior $p(\mathbf{x}_k|\mathcal{Y})$ is then given as a combination of the two messages

$$p(\mathbf{x}_k|\mathcal{Y}) \propto \alpha_k(\mathbf{x}_k)\beta_k(\mathbf{x}_k). \tag{2.43}$$

For linear Gaussian systems both $\alpha_k(\mathbf{x}_k)$ and $\beta_k(\mathbf{x}_k)$ are Gaussian so that the product $p(\mathbf{x}_k|\mathcal{Y})$ is also Gaussian. For most problems, including those considered here, such closed forms solutions do not exist, so that approximations must be used. When compared to the standard forward-backward approach, the two-filter approach is advantageous as it allows for parallel implementation. More importantly, it is required for deriving tractable computational updates when computing recursions in a VBEM framework; see Beal (2003) Sect. 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 Θ are also required to be estimated, a joint field-parameter estimation algorithm is required. One such methods is the focus of the following sub-sections.

2.3.3 The VBEM Algorithm

Parameter estimation in latent variable models (of which linear dynamical systems are an example) is frequently carried out using the *expectation-maximization* (EM) algorithm (Dempster et al. 1977) an elegant iterative algorithm which is guaranteed to converge to a local optimum of the marginal likelihood function (i.e. the joint where the states have been marginalized). However, EM does not provide an estimate of the posterior uncertainty over parameters; while heuristics have been proposed, proper quantification of the uncertainty over parameters can only be obtained through a computation of the posterior distribution in a Bayesian framework. Unfortunately, in general such a computation is impossible. To see why, consider the simple case of two Gaussian random variables whose product is observed with Gaussian noise,

$$x = ab + \varepsilon \quad \varepsilon \sim \mathcal{N}(0, \sigma^2), \quad a, b \sim \mathcal{N}(0, 1).$$

One variable (say a) can be easily integrated out since both $p(x|a, b)$ and $p(a)$ are Gaussian. The result of this marginalization yields $p(x|b)$ which is Gaussian in x but with terms in b^2 in the *denominator* of the exponent. This makes marginalization of b with a Gaussian prior, required to obtain the marginal likelihood $p(x)$, not possible in closed form. This example, while simple, is representative of the problems encountered in linear dynamical systems, where the state variable is multiplied by a parameter at every time point to give the temporal evolution.

Computation of the joint posterior over states and parameters can thus only be done approximately. Broadly speaking, there are two classes of approximate inference algorithms which are widely used in the statistical and machine learning communities. *Sampling algorithms* exploit the fact that, while it is impossible to compute the posterior distribution, it is often possible to produce an algorithm which will eventually generate samples from the posterior (typically through the construction of a Markov chain which will converge to the distribution asymptotically). While these methods have come to be regarded as a gold standard as they are provably exact

asymptotically, assessing convergence is difficult, and the inherent scaling of the sampling error (which decreases inversely to the square root of the number of samples) means that obtaining accurate results is always computationally demanding. Here, we use the *variational* approach to inference, and in particular VBEM.

VBEM is a framework for analytic computations of approximate posterior distributions over latent variables and parameters, proposed in Attias (1999, 2000) to solve the state and parameter estimation problem in linear dynamical systems. The posterior distributions are computed using iterations (coined Iterative VB in Šmídl and Quinn (2005), Sect.1.2), in a similar way as the EM algorithm, and its convergence is guaranteed. Whilst inheriting the advantages of being a Bayesian approach and thus establishing credible intervals over an unknown set of parameters, the method is deterministic, i.e. no sampling is required rendering it (generally) faster than standard Markov chain Monte Carlo (MCMC) approaches. VB has seen applicability in a wide range of problems such as the modeling of the cell's regulatory network (Beal et al. 2005; Sanguinetti et al. 2006) and vision tracking (Vermaak et al. 2003).

The VB method hinges on the definition of an objective functional which makes the inference problem equivalent to an optimization problem. The natural choice is the *Kullback-Leibler (KL) divergence*, and information theoretic quantity which determines the cross-entropy between two distributions

$$KL[q \| p] = \int dq \log \frac{q}{p}. \quad (2.44)$$

It can easily be proved that the KL divergence is a convex functional of its first argument q , and that it is zero iff $q = p$ in distribution (Cover and Thomas 2012). If we now let p be the posterior distribution of our system (defined through Bayes' theorem), minimizing the KL divergence with respect to q will therefore yield the correct posterior. Naturally, this is just as intractable as computing the posterior ab initio; the key to the success of variational methods lies in finding a convenient functional form for the approximating distribution q . Notice that the integrals in equation (2.44) usually involve computations of moments of the approximating distribution q , so that a convenient choice can yield analytic solutions. In the VBEM method, the approximation is carried out using conditionally independent distributions $\tilde{p}(\mathcal{X})$ and $\tilde{p}(\Theta)$ so that $q = \tilde{p}(\mathcal{X})\tilde{p}(\Theta)$. Throughout this work $\tilde{p}(\mathcal{X})$ and $\tilde{p}(\Theta)$ will be referred to as the *variational posterior distributions*. The forms of these distributions are obtained by minimizing the KL divergence (2.44) by setting to zero its functional derivatives. It can be shown that this condition is met for

$$\tilde{p}(\mathcal{X}) \propto \exp \left(\mathbb{E}_{\tilde{p}(\Theta)} [\ln p(\mathcal{X}, \Theta, \mathcal{Y})] \right), \quad (2.45)$$

$$\tilde{p}(\Theta) \propto \exp \left(\mathbb{E}_{\tilde{p}(\mathcal{X})} [\ln p(\mathcal{X}, \Theta, \mathcal{Y})] \right), \quad (2.46)$$

where $\mathbb{E}_{\tilde{p}(\cdot)}[\cdot]$ is used to render specific the distribution relative to which we are taking expectations. Notice that while \mathcal{X} and Θ are independent random variables

under the approximating distribution q , each of their marginal distributions *depends on the statistics* of the other.

Due to the inter-dependence between the variational posteriors, (2.45) and (2.46) cannot be solved directly. An iterative algorithm, the VBEM algorithm, is thus required. This operates by (i) considering a parameter distribution $\tilde{p}(\boldsymbol{\Theta})^{(i)}$ and finding $\tilde{p}(\mathcal{X})^{(i+1)}$ (ii) fixing $\tilde{p}(\mathcal{X})^{(i+1)}$ and finding $\tilde{p}(\boldsymbol{\Theta})^{(i+1)}$ and (iii) re-iterating until convergence to a (local) maximum is reached, see Fig. 2.4. Convergence may be assessed by monitoring the change in the objective function across consecutive iterations. However, frequently, other quantities, which are readily computed, are monitored, such as the mean of the parameter posterior distribution. A summary of the VBEM algorithm is given in Algorithm 2.1.

Algorithm 2.1: The VBEM algorithm

Input: Data set \mathcal{Y} , initial parameter variational posterior distribution $\tilde{p}(\boldsymbol{\Theta})^{(0)}$.

$i = 0$

while (not converged)

$$\tilde{p}(\mathcal{X})^{(i+1)} \propto \exp \left(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta})^{(i)}} [\ln p(\mathcal{X}, \boldsymbol{\Theta}, \mathcal{Y})] \right) \quad \text{VBE-step}$$

$$\tilde{p}(\boldsymbol{\Theta})^{(i+1)} \propto \exp \left(\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}} [\ln p(\mathcal{X}, \boldsymbol{\Theta}, \mathcal{Y})] \right) \quad \text{VBM-step}$$

$i = i + 1$

Output: $\tilde{p}(\mathcal{X})^{(i)}$, $\tilde{p}(\boldsymbol{\Theta})^{(i)}$.

VBEM exhibits many similarities to 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 maximum likelihood 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 (Zammit Mangion et al. 2011b).

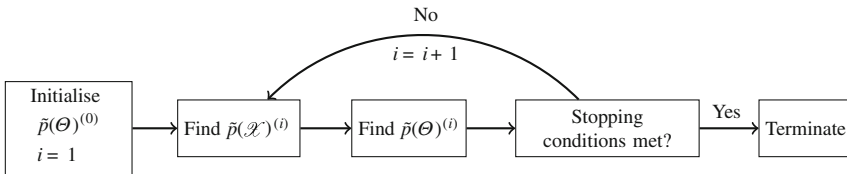


Fig. 2.4 Flow-chart representation of the VBEM algorithm

2.4 Implementation Tools

Sections 2.1–2.3 have given an overview of the theory on spatio-temporal modeling including details on state/parameter estimation. The analyst has to make several choices when ensuring parsimony of the adopted model. In particular, what set of basis functions should be employed? And how can the model space of the SIDE be reduced to one able to reconstruct observed data? The answers to both these questions lie in the use of non-parametric tools for providing descriptive insights into the data.

2.4.1 Non-Parametric Description of Point Patterns

Section 2.1.2 introduced the LGCP as a conventional approach to modeling double stochasticity in point processes. LGCPs are commonly employed because they have analytic properties which are readily exploited in modeling and estimation. To see this consider a Gaussian random variable $x \sim \mathcal{N}(\mu, \sigma^2)$ and set $y = e^x$. Then, $\mathbb{E}[y] = \exp(\mu + \sigma^2/2)$ and $\mathbb{E}[y^2]/\mathbb{E}[y]^2 = \exp(\sigma^2)$. Hence, by computing the empirical moments of y , one can obtain an estimate of the parameters describing the distribution of x . This is the principle behind the use of non-parametric fits in LGCPs.

In fact the log-Gaussian properties for a univariate Gaussian variable easily extend to the spatio-temporal case. Consider, once again, a log-Gaussian intensity process $\lambda_k(s) = \exp(z_k(s))$ within a single time-frame k . Let $\text{cov}(z_k(s)) = \sigma_k^2 \psi_k(v)$, then the two following properties hold (Møller et al. (1998)):

$$\lambda_k^{(1)}(s) = \exp(\mu_k(s) + \sigma_k^2/2) \quad (2.47)$$

$$\frac{\lambda_{k,k}^{(2)}(s, \mathbf{r})}{\lambda_k^{(1)}(s)\lambda_k^{(1)}(\mathbf{r})} = \exp(\sigma_k^2 \psi_k(s, \mathbf{r})) = g_{k,k}(s, \mathbf{r}), \quad (2.48)$$

where $\lambda_k^{(1)} = \mathbb{E}[\lambda_k(s)]$ and $\lambda_k^{(2)} = \mathbb{E}[\lambda_k(s)\lambda_k(\mathbf{r})]$. The intuition is that if the first- and second-order moments of the intensity function are found, then $\mu_k(s)$ and $\sigma^2 \psi(s, \mathbf{r})$ may be estimated. The quantity $\lambda_k^{(1)}$ is the mean intensity function and $g_{k,k}(s, \mathbf{r})$ is known as the pair auto-correlation function (PACF). The latter quantity relates to the probability of finding a point at \mathbf{r} given that a point is present at s and can reveal several interesting characteristics of a point pattern. In particular, if $g_{k,k}(s, \mathbf{r})$ is flat for a given s then the pattern is said to be entirely random at s ; if g decays as $\|s - \mathbf{r}\|$ increases then clustering is observed, if $\|s - \mathbf{r}\|$ oscillates then the point pattern is regular and so on (see Stoyan and Stoyan 1994, Chap. 15). The PACF is very important in this setting because it gives an indication of the spectral properties of the latent field generating the point process. This is of use in basis function decomposition, described in Sect. 2.4.2.

Another important point is that in spatio-temporal systems one may also consider the pair cross-correlation function (PCCF) which, analogous to (2.48), is given by

$$\frac{\lambda_{k,k+1}^{(2)}(\mathbf{s}, \mathbf{r})}{\lambda_k^{(1)}(\mathbf{s})\lambda_{k+1}^{(1)}(\mathbf{r})} = g_{k,k+1}(\mathbf{s}, \mathbf{r}). \quad (2.49)$$

Now $g_{k,k+1}(\mathbf{s}, \mathbf{r})$ relates to the probability of observing an event at $(k+1, \mathbf{r})$ given that one has observed an event at (k, \mathbf{s}) (Brix and Møller 2001). Clearly this descriptor is of considerable use in conflict scenarios to describe spatio-temporal evolution of the intensity. A PCCF which is a spike centered at \mathbf{s} , for each \mathbf{s} , is indicative of low mobility/interaction, and corresponds to an SIDE with a kernel which is narrow relative to ψ_k . An example of this is given in Chap. 3.

Estimates of the PACF and PCCF rely on non-parametric analysis of the point patterns. In the following we describe those most commonly employed in practice.

Estimation of $\lambda^{(1)}(\mathbf{s})$: If \mathcal{Y}_k is first-order stationary (i.e. $\lambda_k^{(1)}(\mathbf{s}) = \lambda_k^{(1)}$), then an estimator for $\lambda_k^{(1)}$ is given as (Stoyan and Stoyan 1994, Chap. 15)

$$\lambda_k^{(1)} = \frac{N_k}{|\mathcal{O}|}, \quad (2.50)$$

where N_k is the cardinality of \mathcal{Y}_k (i.e. the number of observed points) and $|\mathcal{O}|$ is the domain area. In some cases, this assumption does not hold and one may employ explanatory variables to mark out clear intensity trends (Brix and Diggle 2001) or a non-parametric kernel estimator (Diggle 1985; Møller and Waagepetersen 2004, Sect. 4.3)

$$\lambda_k^{(1)}(\mathbf{s}) = \sum_{s_i \in \mathcal{Y}} \frac{k_b(\|\mathbf{s} - \mathbf{s}_i\|)}{c_{\mathcal{O},b}(\mathbf{s}_i)}. \quad (2.51)$$

Here, $c_{\mathcal{O},b}(\mathbf{s}_i)$ is an edge-correction factor given as $c_{\mathcal{O},b}(\mathbf{s}_i) = \int_{\mathcal{O}} k_b(\|\mathbf{s} - \mathbf{s}_i\|) d\mathbf{s}$ and $k_b(\mathbf{s})$ is a smoothing kernel, the most common of which is the *Epanečnikov kernel*. Equation (2.51) is also very useful as a *visualization* tool of the point pattern and indeed has been used extensively in this way with the WikiLeaks Afghan War Diary; see for instance O’Loughlin et al. (2010b), the animation by Dewar (2010), or McCormick et al. (2010), which shows the intensity of improvised explosive device (IED) attacks which caused casualties. The mean function however gives no indication as to how conflict intensity is correlated in space, this is achieved through use of the PACF $g_{k,k}(\mathbf{s}, \mathbf{r})$.

Estimation of $g_{k,k}(\mathbf{s}, \mathbf{r})$: A common assumption, which we adhere to in this book for preliminary analysis, is second-order stationarity, i.e. $g_{k,k}(\mathbf{s}, \mathbf{r}) = g_{k,k}(\|\mathbf{s} - \mathbf{r}\|) = g_{k,k}(v)$. A non-parametric estimator (on \mathbb{R}^2) for the homogeneous PACF is given by Brix and Diggle (2001); Baddeley et al. (2000)

$$\hat{g}_{k,k}(\nu) = \frac{1}{2\pi\nu|\mathcal{O}|} \sum_{\substack{\neq \\ s_i, s_j \in \mathcal{O}_k}} \frac{k_b(\|s_i - s_j\| - \nu)}{\lambda_k^{(1)}(s_i)\lambda_k^{(1)}(s_j)w(s_i, s_j)}, \quad (2.52)$$

where $w(s_i, s_j)$ is the fraction of the circle (in 2 dimensions) with center s_i and radius $\|s_i - s_j\|$ lying in \mathcal{O} . Intuition into (2.52) can be obtained as follows. Let $\lambda_k^{(1)}(s) = \lambda_k^{(1)}$ be constant over space and $w(s_i, s_j) = 1$ (we are sufficiently within the domain's boundary) and $k_b(r) = [1/(\pi b^2)]\mathbf{1}(r \leq b)$ be a uniform kernel. For $b \ll \nu$ then $\hat{g}_{k,k}(\nu) \propto (\text{no. of pairs a distance } \nu \text{ apart})/\nu$. In a pure Poisson process, the number of pairs grows linearly with ν (since the perimeter of a circle grows linearly in ν) and hence $\hat{g}_{k,k}(\nu)$ is constant. On the other hand in the presence of clustering the number of pairs is high at small ν and grows at a much slower rate, causing a decay in $\hat{g}_{k,k}(\nu)$. This rate of decay is indicative of the cluster size. An estimator (and interpretation) for the PCCF (again on \mathbb{R}^2) follows analogously

$$\hat{g}_{k,k+1}(\nu) = \frac{1}{2\pi\nu|\mathcal{O}|} \sum_{\substack{\neq \\ s_i \in \mathcal{O}_k \\ s_j \in \mathcal{O}_{k+1}}} \frac{k_b(\|s_i - s_j\| - \nu)}{\lambda_k^{(1)}(s_i)\lambda_{k+1}^{(1)}(s_j)w(s_i, s_j)}. \quad (2.53)$$

Once the PACF is known, a parametric representation for the auto-covariance function $\sigma^2\psi_k(s)$ may be readily found from (2.52) using a moment-based method. Throughout this book we will assume that $\psi_k(\|s - \mathbf{r}\|) = \psi_k(\nu)$ is Gaussian and of the form $\psi_k(\nu) = \exp(-\nu^2/2\sigma_b^2)$. Then estimates for σ_k^2 and σ_b^2 are obtained by minimizing

$$\int_{\nu_1}^{\nu_2} \left[\ln \hat{g}_{k,k}(\nu) - \sigma_k^2 \exp\left(-\nu^2/2\sigma_b^2\right) \right]^2 d\nu, \quad (2.54)$$

for some user-defined ν_1, ν_2 . The same procedure is used for estimating the cross-covariance function.

The decaying characteristic of $g_{k,k}(\nu)$ leads to the idea that although spatial patterns might exhibit some correlation, there is a *cutoff* length scale beyond which clusters may be considered to be independent—this is useful for defining the spectral characteristics of a point process, discussed next.

2.4.2 Basis Selection from Point-Process Observations

A requirement for the Galerkin method of Sect. 2.2.3 is that the basis set \mathcal{B} is linearly independent. Further, in practice, this set should be able to accurately approximate (2.17) (Harrington 1993). Clearly a vast range of continuous-space functions may be employed, however some emerge as more useful than others. A popular set in spatio-temporal statistics are the empirical orthogonal functions (EOFs), which describe the prominent spatial features of the evolving fields (e.g. Wikle and Cressie 1999;

Berliner et al. 2000). EOFs are optimal in the sense that the variance of the error due to truncation in (2.17) is minimized. A key hindrance to the use of EOFs in conflict applications, and indeed spatio-temporal point processes in general, is their construction. The EOFs would have to be constructed from an intensity field estimated using non-parametric methods—estimating the intensity is usually the target of the analysis in the first place and it is unclear how the pre-smoothing would affect the adopted basis and hence the final predictions (see Cressie and Wikle 2011, Sect. 7 for related issues with Gaussian data).

We hence suggest the use of functions of local scope, such as finite elements (Lindgren et al. 2011), *Epanečnikov kernels* or Gaussian radial basis functions (GRBFs) (Stroud et al. 2001; Dewar et al. 2009; Scerri et al. 2009). GRBFs have gained interest recently due to analytical solutions readily available for Gaussian functions. In addition, they are universal approximators (Park and Sandberg 1991) and also have well-defined frequency-response functions. In particular, GRBFs have the favorable property that their Fourier transform are also Gaussian in the frequency domain so that

$$\phi(v) = \mathcal{F}\{\phi(s)\} = \sqrt{2\pi\sigma_b^2} \exp(-2\pi^2\sigma_b^2 v^2). \quad (2.55)$$

The variances in the spatial and frequency domain are then related through the mappings (Sanner and Slotine 1992)

$$\sigma_v^2 \leftarrow \frac{1}{4\pi^2\sigma_b^2} \quad \sigma_b^2 \leftarrow \frac{1}{4\pi^2\sigma_v^2}. \quad (2.56)$$

Consider now an LGCP, where the latent GP has covariance function $k(v) = \sigma^2\psi(v)$ and where $\psi(\cdot)$ is from the Matérn class

$$k(s, r) = \frac{1}{2^{\tilde{\nu}-1}\Gamma(\tilde{\nu})} (\kappa||s - r||) K_{\tilde{\nu}}(\kappa||s - r||), \quad (2.57)$$

where $K_{\tilde{\nu}}$ is the modified Bessel function of the second kind and $\tilde{\nu}$ is a smoothness parameter. κ is a length scale parameter such that at distances $\rho = \sqrt{8\tilde{\nu}}/\kappa$, the correlation between two spatial points is approximately 0.1 (Lindgren et al. 2011). ρ is termed the *range parameter*. Since $\psi(v)$ is an auto-correlation function, it is related to the spectral properties of the Gaussian field through the *auto-correlation theorem* (Bracewell 2000, p. 122) which states that the Fourier transform of $\psi(v)$, $\mathcal{F}\{\psi(v)\}$, is the signal's power spectrum:

$$\mathcal{F}\{\psi(v)\} = |Z(v)|^2, \quad (2.58)$$

where $Z(v)$ is the Fourier transform of z . From the spectrum, a cutoff frequency v_c may be identified beyond which higher frequencies can be ignored. v_c is largely a design choice; the 3dB point, defined as the frequency at which the signal is at half-power, is a common one (Freestone et al. 2011). v_c is generally indicative of the

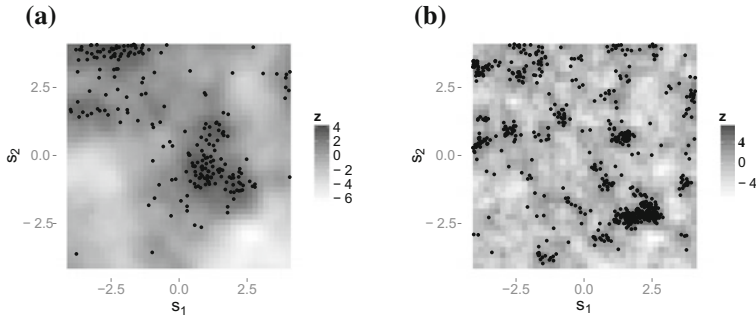


Fig. 2.5 Effect of cutoff frequency/range parameter of latent field (*background*) on cluster size of observed events (*dots*). **a** realization of an LGCP equipped with a $\tilde{\nu} = 3/2$ Matérn kernel with $\rho = 3.5$ units. **b** realization of an LGCP equipped with a $\tilde{\nu} = 3/2$ Matérn kernel with $\rho = 0.7$ units

nature of the clusters in the data—a large ν_c (indicative of small ρ) is a sign of multiple small, high frequency clusters. A low ν_c (indicative of large ρ) is representative of the occurrence of few, broad, clusters, see Fig. 2.5.

Once ν_c is found, we place GRBFs regularly spaced within the spatial domain of interest. Denote these grid points as $\{\xi_i\}_{i=1}^n$. The grid has to be sufficiently fine so as to avoid *aliasing* by satisfying Shannon's sampling criterion (Scerri et al. 2009) i.e. if the grid spacing is Δ_s then

$$\Delta_s < \frac{1}{2\nu_c} = \frac{1}{2\alpha_0\nu_c}, \quad (2.59)$$

where ν_c is the selected frequency cutoff and $\alpha_0 > 1$ is an oversampling parameter.

Second, the frequency bandwidth of the construction basis has to be larger than that of the field. When using GRBFs, Sanner and Slotine (1992) suggest that

$$\sigma_v = \frac{1}{\sqrt{2}}\nu_c. \quad (2.60)$$

Due to the Fourier duality of GRBFs, the basis function width in the spatial domain may be directly specified. By substitution of (2.60) in (2.56)

$$\sigma_b = \sqrt{\frac{1}{2\nu_c^2\pi^2}}. \quad (2.61)$$

The procedure may be summarized as follows. First, estimate the PACF using (2.52) and use this to obtain a Gaussian $\psi(\nu)$ using the optimization routine in (2.54). Second, find the spectral response by computing the Fourier transform of the auto-covariance function in (2.58). Third, arrange the basis on a grid with separation governed by (2.59). Finally, estimate the cutoff frequency and use this to find the

parameter σ_b from (2.61). Note that for a spatio-temporal process, estimation of $\psi(v)$ can be done for each time-frame. An average over all time frames is then taken to obtain a robust spectral estimate.

2.4.2.1 ‘Compact’ GRBFs

GRBFs are of global scope, and hence do not vanish on the boundary of \mathcal{O} , $\partial\mathcal{O}$, a desirable feature in some spatio-temporal applications (for instance those involving Dirichlet boundary conditions). A very similar function to the GRBF, which is of compact support, is given by

$$\phi(s) = \begin{cases} \frac{(2\pi - \|\tau s\|)(1 + (\cos \|\tau s\|)/2) + \frac{3}{2} \sin(\|\tau s\|)}{3\pi}, & \|\tau s\| < 2\pi, \\ 0, & \text{otherwise,} \end{cases} \quad (2.62)$$

for $\tau > 0$ and where $\|\cdot\|$ denotes the usual Euclidean distance on \mathcal{O} . Since (2.62) defines a function of compact support, it is sometimes used to enforce independence between spatial points which are considerably separated in space (Storkey 1999).

The function $\phi(s)$ in (2.62), which will be termed the ‘compact’ GRBF throughout the book, closely resembles the isotropic GRBF with $\phi(s) = \exp(-\tau^2 \|s\|^2 / 2\pi)$. The similarity allows us to assume GRBFs when setting up the basis (i.e. find σ_b using ν_c in (2.61)) and then evaluate τ through

$$\tau = \sqrt{\pi}/\sigma_b = \sqrt{2\nu_c^2\pi^3}. \quad (2.63)$$

Hence, if a compact basis is required, one may instead place compact GRBFs with parameter τ in the spatial domain centered on the coordinates $\{\xi_i\}_{i=1}^n$.

2.4.3 Approximate Inference from Point Observations

As mentioned earlier, the non-Gaussian nature of the observation model (2.33) leads to intractable integrals in the recursive computation of the forward and backward messages of (2.39) and (2.42). Hence, exact computations using the two filter smoother are not possible. For this reason we need to introduce approximations in the message passing algorithm—one such approximation is the Laplace approximation. A full account of how this fits in with the variational two-filter smoother is given in Sect. 2.4.4. Here, we show what computations it entails by applying it to a spatial point process.

Consider a spatial conditionally Poisson point process \mathcal{Y} with the likelihood (2.33). With \mathbf{x} equipped with a Gaussian prior, the posterior $p(\mathbf{x}|\mathcal{Y})$ is intractable. The density however may be approximated to a Gaussian using a Laplace approximation (Bishop 2006, Sect. 4.4). This Gaussian density is centered at the mode

of the posterior density and has a precision matrix equal to the negative Hessian of $p(\mathbf{x}|\mathcal{Y})$ at the mode. In the spatial point process, the mode and Hessian are found from (2.33) as follows. First, let \mathbf{x} have a prior mean \mathbf{x}_p and variance Σ . The mode $\hat{\mathbf{x}}$ is then the point at which

$$\left. \frac{\partial}{\partial \mathbf{x}} \ln p(\mathbf{x}|\mathcal{Y}) \right|_{\mathbf{x}=\hat{\mathbf{x}}} = \mathbf{0}, \quad (2.64)$$

where

$$\frac{\partial}{\partial \mathbf{x}} \ln p(\mathbf{x}|\mathcal{Y}) = \sum_{s_j \in \mathcal{Y}} \boldsymbol{\phi}(s_j) - \int_{\mathcal{O}} \boldsymbol{\phi}(s) \exp(\boldsymbol{\phi}(s)^T \mathbf{x}) ds - \Sigma^{-1}(\mathbf{x} - \mathbf{x}_p). \quad (2.65)$$

This optimization may be carried out using a gradient ascent method such as conjugate gradient or its scaled version. The integral within the optimization may be efficiently computed using numerical quadrature since $\mathcal{O} \in \mathbb{R}^2$ and, for the purposes of this book, the number of basis functions n is small. With some types of basis functions, such as finite elements, this integral may be computed accurately and quickly even for large n , see for instance Simpson et al. (2011).

The Hessian is given by

$$\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}^T} \ln p(\mathcal{Y}|\mathbf{x}) = - \int_{\mathcal{O}} \boldsymbol{\phi}(s) \boldsymbol{\phi}(s)^T \exp(\boldsymbol{\phi}(s)^T \mathbf{x}) ds - \Sigma^{-1}. \quad (2.66)$$

and

$$\hat{\Sigma} = \text{cov}(\mathbf{x}|\mathcal{Y}) \approx - \left[\left. \frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}^T} \ln p(\mathcal{Y}|\mathbf{x}) \right|_{\mathbf{x}=\hat{\mathbf{x}}} \right]^{-1}, \quad (2.67)$$

which can be easily computed for small n .

In the context of point processes, the Laplace approximation was first used within an expectation maximization (EM) framework in Smith and Brown (2003). The advantage of the approximate inference method is that, in conjunction with dimensionality reduction, it is generally quick and allows for prediction with ease due to normality assumptions. In particular, if the distribution of the latent field at \mathbf{s}^* , $z(\mathbf{s}^*)$ is desired, then this is simply given as

$$z(\mathbf{s}^*) \sim \mathcal{N}(\boldsymbol{\phi}(\mathbf{s}^*)^T \hat{\mathbf{x}}, \boldsymbol{\phi}(\mathbf{s}^*)^T \hat{\Sigma} \boldsymbol{\phi}(\mathbf{s}^*)), \quad (2.68)$$

where $p(\mathbf{x}|\mathcal{Y}) \simeq \mathcal{N}(\hat{\mathbf{x}}, \hat{\Sigma})$ is the approximate posterior density of \mathbf{x} . For several applications, the Laplace approximation will be sufficient. However, in Zammit Mangion et al. (2011b) it was seen that improvements might be made when employing the Laplace method within a variational Bayes (VB) framework for inference in a spatio-temporal setting. These ideas are formalized in a VB-Laplace framework.

2.4.4 VB-Laplace Inference from Point-Process Observations

Laplace approximations are used to keep the variational updates/recursions in Algorithm 2.1 tractable. We term the combination of Laplace and variational Bayes VB-Laplace. In this section we focus on the state update equations and assume that conditioned on normally distributed states, the parameter updates remain tractable. The latter updates will be shown for the specific case of the Afghan War Diary in Appendix A.

Assume that the spatio-temporal system governed by an SIDE or SPDE has been reduced into state-space form using the basis function placement methods described in Sect. 2.4.2. Then, the model under study is

$$\mathbf{x}_{k+1} = \mathbf{A}(\boldsymbol{\theta})\mathbf{x}_k + \mathbf{w}_k(\boldsymbol{\theta}), \quad (2.69)$$

$$\lambda_k = f(\mathbf{b}) \exp(\boldsymbol{\phi}(\mathbf{s})^T \mathbf{x}_k), \quad (2.70)$$

where $\boldsymbol{\theta}$ are unknown parameters in the evolution equation to be estimated and where we have introduced $f(\mathbf{b})$ as an intensity component which encodes covariate information. Typically $f(\mathbf{b})$ is log-linear with geo-demographic features, such as population density or elevation as weighted covariates. The weights, \mathbf{b} , are unknown and also need to be estimated (recall 2.37). Let $\boldsymbol{\Theta} = [\boldsymbol{\theta}^T, \mathbf{b}^T]^T$. The task here is to find an approximate solution to the VB recursions which we re-express here for completeness as

$$\tilde{p}(\mathcal{X})^{(i+1)} \propto \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta})^{(i)}} [\ln p(\mathcal{X}, \mathcal{Y}, \boldsymbol{\Theta})]), \quad (2.71)$$

$$\tilde{p}(\boldsymbol{\Theta})^{(i+1)} \propto \exp(\mathbb{E}_{\tilde{p}(\mathcal{X})^{(i+1)}} [\ln p(\mathcal{X}, \mathcal{Y}, \boldsymbol{\Theta})]), \quad (2.72)$$

where, for each iteration, $\tilde{p}(\mathcal{X})$ is found using a two-filter smoother. It can be shown that under VB approximations, the forward message (2.39) becomes (Beal 2003)

$$\tilde{\alpha}_k(\mathbf{x}_k) \propto \int \tilde{\alpha}_{k-1}(\mathbf{x}_{k-1}) \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta})^{(i)}} [\ln p(\mathbf{x}_k | \mathbf{x}_{k-1}, \boldsymbol{\Theta}) p(\mathbf{y}_k | \mathbf{x}_k, \boldsymbol{\Theta})]) d\mathbf{x}_{k-1}.$$

The marginalization (i.e. propagation) of \mathbf{x}_{k-1} is carried out analytically. Laplace approximations are then used to combine the propagated density with the point-process likelihood as follows

$$\tilde{\alpha}_k(\mathbf{x}_k) \propto \mathcal{N}_{\mathbf{x}_k}(\tilde{\mathbf{x}}_k, \tilde{\Sigma}_k) \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta})^{(i)}} [\ln p(\mathbf{y}_k | \mathbf{x}_k, \boldsymbol{\Theta})]) \xrightarrow{\text{Laplace}} \mathcal{N}_{\mathbf{x}_k}(\hat{\mathbf{x}}_{k|k}, \Sigma_{k|k}). \quad (2.73)$$

Approximating the VB forward message to a Gaussian assures that the recursions are maintained. The same is required for the backward message where

$$\begin{aligned} \tilde{\beta}_k(\mathbf{x}_k) &= \int \tilde{\beta}_{k+1}(\mathbf{x}_{k+1}) \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta})}[\ln p(\mathbf{x}_{k+1}|\mathbf{x}_k, \boldsymbol{\Theta}) p(\mathbf{y}_{k+1}|\mathbf{x}_{k+1}, \boldsymbol{\Theta})]) d\mathbf{x}_{k+1} \\ &\xrightarrow{\text{Laplace}} \mathcal{N}_{\mathbf{x}_k}(\hat{\mathbf{x}}_{k|k+1:K}, \Sigma_{k|k+1:K}). \end{aligned} \quad (2.74)$$

The two messages are then combined to give the smoothed estimate:

$$\begin{aligned} \tilde{p}(\mathbf{x}_k | \mathbf{y}_{1:K}) &\propto \tilde{p}(\mathbf{x}_k | \mathbf{y}_{1:k}) \tilde{p}(\mathbf{y}_{k+1:K} | \mathbf{x}_k) \\ &= \tilde{\alpha}_k(\mathbf{x}_k) \tilde{\beta}_k(\mathbf{x}_k) = \mathcal{N}_{\mathbf{x}_k}(\hat{\mathbf{x}}_{k|K}, \Sigma_{k|K}). \end{aligned} \quad (2.75)$$

In addition to the marginal variances computed in (2.75), in spatio-temporal systems the cross-covariance matrix describing the interactions across time and space is also required. This is obtained through the joint (Beal 2003, Sect. 5.3.5)

$$\tilde{p}(\mathbf{x}_k, \mathbf{x}_{k-1} | \mathcal{Y}) = \tilde{\alpha}_{k-1}(\mathbf{x}_{k-1}) \tilde{\beta}_k(\mathbf{x}_k) \exp(\mathbb{E}_{\tilde{p}(\boldsymbol{\Theta})^{(i)}}[p(\mathbf{x}_k | \mathbf{x}_{k-1}, \boldsymbol{\Theta}) p(\mathbf{y}_k | \mathbf{x}_k, \boldsymbol{\Theta})]), \quad (2.76)$$

which also needs to be approximated using a Laplace approximation. This involves the inverse of a 2×2 block precision matrix which may be carried out using Schur complements (Minka 2000; Beal 2003). In the interest of brevity we omit computational details and refer the reader to (Beal 2003, Chap. 5).

2.5 Conclusion

This chapter developed the concepts first presented in Chap. 1 into a unified framework for studying spatio-temporal point-process systems with application to conflict. Several alterations to the IDE-Cox (or SPDE-Cox) combination may be carried out. For instance nonlinear spatio-temporal models might be utilized (e.g. Freestone et al. 2011) or the intensity function may be altered to cater for self-exciting events (and thus causality) resulting in a Hawkes-Cox process likelihood (Mohler 2013). Such considerations complicate modeling and inference to a certain degree, but are nonetheless readily accommodated in the described framework which in itself is very flexible to different scenarios. Next, in Chap. 3, we show how the techniques elucidated above are ideally placed for modeling, estimation and prediction in conflict.

References

- Anderson BDO, Moore J, Barratt J (1979) Optimal filtering. Prentice-Hall, New Jersey
- Attias H (1999) Inferring parameters and structure of latent variable models by variational Bayes. In: Proceedings of the 15th conference on uncertainty in artificial intelligence, pp 21–30
- Attias H (2000) A variational Bayesian framework for graphical models. In: Advances in neural information processing systems, vol 12. pp 209–215
- Baddeley AJ, Møller J, Waagepetersen R (2000) Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. Stat Neerl 54(3):329–350

- Beal MJ (2003) Variational algorithms for approximate Bayesian inference. PhD thesis, University College London, UK
- Beal MJ, Falciani F, Ghahramani Z, Rangel C, Wild DL (2005) A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics* 21(3):349
- Berliner LM, Wikle CK, Cressie N (2000) Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *J Climate* 13(22):3953–3968
- Bishop CM (2006) Pattern recognition and machine learning. Springer, New York
- Bracewell RN (2000) The Fourier transform & its applications, 3rd edn. McGraw-Hill, Singapore
- Briers M, Doucet A, Maskell S (2004) Smoothing algorithms for state-space models. Technical Report, TR-CUED-F-INFENG 498, University of Cambridge
- Brix A, Møller J (2001) Space-time multi type log Gaussian Cox processes with a view to modelling weeds. *Scand J Stat* 28(3):471–488
- Brix A, Diggle PJ (2001) Spatiotemporal prediction for log-Gaussian Cox processes. *J Roy Stat Soc B* 63(4):823–841
- Carmona RA (1998) Stochastic partial differential equations: six perspectives. American Mathematical Society, Providence
- Coleman MP (2005) An introduction to partial differential equations with Matlab. Chapman and Hall/CRC, London
- Cover TM, Thomas JA (2012) Elements of information theory. Wiley & Sons, New York
- Cressie NAC, Wikle CK (2011) Statistics for spatio-temporal data. Wiley, New Jersey
- Cseke B, Heskes T (2011) Approximate marginals in latent Gaussian models. *J Mach Learn Res* 12:417–454
- Cseke B, Zammit-Mangion A, Sanguinetti G, Heskes T (2013) Sparse approximations in spatio-temporal point-process models. <http://arxiv.org/abs/1305.4152v2>. Accessed 08 June 2013
- da Prato G, Zabczyk J (1993) Stochastic equations in infinite dimensions. Cambridge University Press, Cambridge
- Dalang RC, Frangos NE (1998) The stochastic wave equation in two spatial dimensions. *Ann Probab* 26(1):187–212
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc B* 39(1):1–38
- Dewar M (2010) Visualisation of activity in Afghanistan using the Wikileaks data. <http://vimeo.com/14200191>. Accessed 28 June 2013
- Dewar M, Scerri K, Kadirkamanathan V (2009) Data-driven spatio-temporal modeling using the integro-difference equation. *IEEE Trans Sig Proc* 57(1):83–91
- Diggle P (1985) A kernel method for smoothing point process data. *App Stat* 34:138–147
- Evans LC (1998) Partial Differential Equations. Graduate studies in mathematics, vol. 19. American Mathematical Society, Providence, RI
- Freestone DR et al (2011) A data-driven framework for neural field modeling. *NeuroImage* 56(3):1043–1058
- Grossmann C, Roos HG, Stynes M (2007) Numerical treatment of partial differential equations. Springer-Verlag, Berlin
- Harrington RF (1993) Field computation by moments method. IEEE Press, Piscataway
- Hausenblas E (2003) Approximation for semilinear stochastic evolution equations. *Potential Anal* 18(2):141–186
- Jazwinski AH (1970) Stochastic processes and filtering theory. Academic Press, London
- Kingman JFC (1992) Poisson processes, vol 3. Clarendon Press, Oxford
- Kot M, Schaffer WM (1986) Discrete-time growth-dispersal models. *Math Biosci* 80(1):109–136
- Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J Roy Stat Soc B* 73(4):423–498
- McCormick M, Allen P, Dant A (2010) 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>. Accessed 30 June 2013

- Minka T (2000) Old and new matrix algebra useful for statistics. <http://research.microsoft.com/~minka/papers/matrix/>. Accessed 30 June 2013
- Mohler G (2013) Modeling and estimation of multi-source clustering in crime and security data. *Ann App Stat* Accepted for publication
- Møller J, Syversveen AR, Waagepetersen RP (1998) Log Gaussian Cox processes. *Scand J Stat* 25(3):451–482
- Møller J, Waagepetersen RP (2004) Statistical inference and simulation for spatial point processes. CRC Press, Boca Raton
- O’Loughlin J, Witmer FDW, Linke AM, Thorwardson N (2010b) Peering into the fog of war: the geography of the Wikileaks Afghanistan war logs, 2004–2009. *Eurasian Geogr Econ* 51(4):472–495
- Park J, Sandberg I (1991) Universal approximation using radial-basis-function networks. *Neural Compu* 3:246–257
- Prévôt C, Röckner M (2007) A concise course on stochastic partial differential equations. Springer-Verlag, Berlin
- Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. MIT Press, Cambridge
- Ross SM (2006) Introduction to probability models. Academic press, London
- Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J Roy Stat Soc B* 71:319–392
- Sanguinetti G, Lawrence ND, Rattray M (2006) Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics* 22(22):2775–2781
- Sanner RM, Slotine JJE (1992) Gaussian networks for direct adaptive control. *IEEE Trans Neural Networ* 3(6):837–863
- Scerri K (2010) A systems approach to spatio-temporal modelling. PhD thesis, University of Sheffield
- Scerri K, Dewar M, Kadirkamanathan V (2009) Estimation and model selection for an IDE-based spatio-temporal model. *IEEE Trans Sig Proc* 57(2):482–492
- Simpson D, Illian J, Lindgren F, Sørbye S, Rue H (2011) Going off grid: Computationally efficient inference for log-Gaussian Cox processes. <http://arxiv.org/abs/1111.0641>. Accessed 08 June 2013
- Šmídl V, Quinn A (2005) The variational bayes method in signal processing. Springer-Verlag, New York
- Smith AC, Brown EN (2003) Estimating a state-space model from point process observations. *Neural Compu* 15(5):965–991
- Storkey AJ (1999) Truncated covariance matrices and Toeplitz methods in Gaussian processes. In: *Proceedings of the international conference on artificial neural networks*, vol 1. pp 55–60
- Stoyan D, Stoyan H (1994) *Fractals, random shapes, and point fields: methods of geometrical statistics*. Wiley, New York
- Stroud JR, Müller P, Sanso B (2001) Dynamic models for spatiotemporal data. *J Roy Stat Soc B* 63:673–689
- Vermaak J, Lawrence N, Perez P (2003) Variational inference for visual tracking. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol 1. pp 773–780
- Wikle C, Cressie N (1999) A dimension-reduced approach to space-time Kalman filtering. *Biometrika* 86(4):815–829
- Wikle CK (2002) A kernel-based spectral model for non-Gaussian spatio-temporal processes. *Stat Model* 2(4):299–314
- Zammit Mangion A, Yuan K, Kadirkamanathan V, Niranjan M, Sanguinetti G (2011b) Online variational inference for state-space models with point-process observations. *Neural Compu* 23(8):1967–1999
- Zammit Mangion A, Sanguinetti G, Kadirkamanathan V (2011a) A variational approach for the online dual estimation of spatiotemporal systems governed by the IDE. In: *Proceedings of the 18th IFAC world congress*, pp 3204–3209

- Zammit-Mangion A, Sanguinetti G, Kadirkamanathan V (2012b) Variational estimation in spatiotemporal systems from continuous and point-process observations. *IEEE Trans Sig Proc* 60(7):3449–3459
- Zammit-Mangion A, Dewar M, Kadirkamanathan V, Sanguinetti G (2012a) Point process modelling of the Afghan War Diary. *P Natl Acad Sci USA* 109(31):12,414–12,419
- Zhukov YM (2012) Roads and the diffusion of insurgent violence: the logistics of conflict in Russia's North Caucasus. *Polit Geogr* 31(3):144–156

Modeling Conflict Dynamics with Spatio-temporal Data

Zammit-Mangion, A.; Dewar, M.; Kadiramanathan, V.;

Flesken, A.; Sanguinetti, G.

2013, VIII, 74 p. 13 illus., 1 illus. in color., Softcover

ISBN: 978-3-319-01037-3