UNIVERSITY OF BATH

University of Bath

Sequential Empirical Bayes Method for Filtering Dynamic Spatiotemporal Processes

Evangelos Evangelou[1]   and Vasileios Maroulas[2]

[1] Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK.
[2] Department of Mathematics, University of Tennessee, Knoxville, TN 37996, USA.

15 June 2017

Address for correspondence: Evangelos Evangelou, Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK. email: ee224@bath.ac.uk

Abstract

We consider online prediction of a latent dynamic spatiotemporal process and estimation of the associated model parameters based on noisy data. The problem is motivated by the analysis of spatial data arriving in real-time and the current parameter estimates and predictions are updated using the new data at a fixed computational cost. Estimation and prediction is performed within an empirical Bayes framework with the aid of Markov chain Monte Carlo samples. Samples for the latent spatial field are generated using a sampling importance resampling algorithm with a skewed-normal proposal and for the temporal parameters using Gibbs sampling with their full conditionals written in terms of sufficient quantities which are updated online. The spatial range parameter is estimated by a novel online implementation of an empirical Bayes method, called herein *sequential empirical Bayes* method. A simulation study shows that our method gives similar results as an offline Bayesian method. We also find that the skewed-normal proposal improves over the traditional Gaussian proposal. The application of our method is demonstrated for online monitoring of radiation after the Fukushima nuclear accident.

Keywords: Dynamic spatiotemporal process; Empirical Bayes estimation; Fukushima nuclear disaster; Geostatistics; Online inference; State space models.

1 Introduction

Many problems related to ecology, epidemiology, defense, and economics exhibit a simultaneous variability in space and time (Cressie and Wikle, 2011; Shaddick and Zidek, 2016). Daily levels of precipitation, temperature, or other environmental variables across a region in a year, e.g. the monitoring of pollutants, the estimation of trajectories of biological entities, or the monitoring of mobile threats within a sensor network (Paci et al., 2013; Ren et al., 2015; Maroulas and Nebenfuhr, 2015; Ren et al., 2016) are a few of a gamut of paradigms which require the careful treatment of spatiotemporal processes in real time.

Our motivation for this study is the monitoring of radiation after the Fukushima nuclear accident. Following the accident, radioactive material was released in the environment which rendered the surrounding areas inhabitable. The authorities established a monitoring program to measure radiation by sampling at specific locations within the infected area. These samples, which are collected daily, can be used to produce a heatmap indicating dangerous zones. This requires spatiotemporal modeling of radiation in real-time to incorporate the new information as soon as it is received. As radiation is measured by the number of nuclear decays emitted each second, a Gaussian model would be inappropriate in this case and a more flexible model is needed.

In cases concerning the study of radiation, disease spreading, weather, and target tracking, where there is no indication when the study will terminate while real-time information must be incorporated to facilitate a rapid-response system, online methods which can assimilate the new data in a fixed computational cost are urgently needed over offline methods. Indeed, the computing costs of offline methods increase commensurately with time and become slower as more and more data are collected. Especially in the modern days of big data, offline methods become inefficient to run for long time periods while online methods are able to take advantage the often-assumed Markov property of the model to simplify computations.

The general setup consists of a latent dynamic process with Markovian evolution while the observation process possesses a conditional independence property. This setup is known as state-space or hidden Markov model. The problem is to compute the filtering distribution, i.e. the distribution of the current value of the latent process given the full observation history up to the present time. A well-known example of an online method is the Kalman filter which gives the

exact filtering distribution in the case of linear Gaussian models. A widely-applicable alternative is particle filtering, or sequential Monte-Carlo, which is a sequential importance sampling method that approximates the filtering distribution by a set of weighted samples (Doucet et al., 2001). On the other hand, it is well known that particle filtering does not perform well when the dimension of the latent process is large, as is the case of many spatiotemporal applications (Cressie and Wikle, 2011, Section 8.4.6).

In addition to the filtering problem, it is also required to estimate the unknown parameters of the model using the observed data on the fly. In a Bayesian setting, this amounts to computing the posterior distribution of these parameters, however the problem now becomes significantly more difficult as there is no straightforward way of integrating out the latent process in an online fashion. A comprehensive review of online estimation methods can be found in Kantas et al. (2015). An obvious solution is to augment the latent process with the parameters, and estimate the joint distribution together by standard particle methods. However, this was recognized by Kitagawa (1998) that this procedure leads to erroneous estimation. An alternative method, proposed by Liu and West (2001), is to impose artificial dynamics on the parameters and estimate them simultaneously with the latent process using particle filters. However, this solution introduces bias in the estimation and requires a significant amount of tuning. The alternative, proposed by Storvik (2002); Fearnhead (2002), is to write the full conditionals of the parameters in terms of sufficient statistics which can be updated sequentially. This approach facilitates sampling from the full conditional of the parameters, however it does not avoid the degeneracy problem and is only applicable to those models where sufficient statistics exist, which is not the case for spatial models where the spatial range parameter is unknown.

This manuscript focuses on the methodology for *online* filtering of a dynamic spatiotemporal process and estimation of the associated static parameters based on data from an exponential family, i.e. the memory and computing time of our method does not grow with time. The inferential procedure derived in this paper is outlined below:

1. For a given spatial range, ϕ, we develop an online algorithm (Algorithm 1) for sampling from the filtering distribution and the posterior distribution of the other parameters. Our algorithm takes advantage of the skewness of the filtering distribution (see Lemma 1) and produces non-degenerate samples.

2. We develop an online estimate of the Bayes factors corresponding to a finite set of ϕ values which we then maximize to estimate ϕ. The asymptotic validity of our method is established in Theorem 1.

3. Given an estimate of ϕ, we derive a novel importance resampling algorithm (Algorithm 2) for estimating the latent process and the other parameters in an online fashion.

The rest of the paper is organized as follows. Section 2 discusses the problem formulation and displays preliminary results related to our model, including the derivation of the skew-normal proposal and the proposed algorithm for sampling for fixed ϕ. Next, Section 3 presents the main contribution of this manuscript and considers the estimation of the spatial correlation via our novel online implementation of the empirical Bayes technique. In Section 4 we illustrate the application of the proposed method for online monitoring of radiation. Section 5 offers a summary and discusses future research directions based on our technique. The proofs to the results presented in the paper are provided in the Appendix. In the Supplementary Materials for this article we present our simulation study for assessing the performance of the proposed method, including a comparison against a typical offline method.

2 Problem formulation and preliminary results

2.1 Model

We adopt a hierarchical, dynamical spatiotemporal model with the observation process Y having conditional independent components from an exponential family distribution given a latent Gaussian spatiotemporal field X.

Let $\mathbb{S}$ denote the continuous spatial domain of interest and let $\mathbb{T} = \{0, 1, 2, \ldots\}$ be the temporal domain. The process model consists of a spatiotemporal Gaussian process X on $\mathbb{S} \times \mathbb{T}$ having an autoregressive structure, that is, for every finite collection of spatial locations in $\mathbb{S}$, say of size n, the value of X at the n locations at time t, $\mathbf{x}_t$, is described by a perturbation of its values at time $t - 1$, $\mathbf{x}_{t-1}$, as follows

$$\mathbf{x}_0 = G_0\beta + \sigma\epsilon_0, \quad \mathbf{x}_t = G_t\beta + \alpha(\mathbf{x}_{t-1} - G_{t-1}\beta) + \sigma\epsilon_t, \ t = 1, 2, \ldots \tag{1}$$

where the driving noise is a normally distributed isotropic spatial process, $\epsilon_t \sim N_n(0, R(\phi))$, G_t defines the $n \times m$ matrix of covariates for each time t associated with an $m \times 1$ parameter vector β, σ is the diffusion coefficient and ϕ denotes the range parameter with spatial correlation matrix $R = R(\phi)$. Our model assumes linear transition in time. In general, α is an $n \times n$ matrix of redistribution weights, however, as Cressie and Wikle (2011) argue in Section 7.2, this can be challenging to estimate at early times. A simpler model that allows stationarity is to let α be a scalar, where $|\alpha| < 1$. This implies homogeneous transition between the components of the state process. The theory developed in this paper applies to either matrix or scalar α but the issues we wish to address can be illustrated more clearly when α is a scalar so we will focus on this case for the remainder of this paper.

The observation process Y is defined on $\mathbb{S} \times \mathbb{T}$ and is assumed to have conditionally independent components given the latent spatiotemporal process X with distribution from an exponential family. In other words, let $y_{i,t}$ denote the value of the observation process at location i and time t, and let $x_{i,t}$ denote the value of the corresponding spatiotemporal process. Then

$$p(y_{i,t}|x_{i,t}) \propto \exp\{y_{i,t}g(x_{i,t}) - \tau_{i,t}b(x_{i,t})\}, \tag{2}$$

where under the usual regularity assumptions for exponential families the mean of the distribution in (2) is $h(x_{i,t})$ with $h(\cdot)$ being the inverse link function, $g(\cdot)$ and $b(\cdot)$ are known functions and $\tau_{i,t}$ is a known scalar associated with the underlying distribution of the data. Further, given $x_{i,t}$, $y_{i,t}$ is independent of every other component of Y. The main advantage in using the general state space representation of a dynamic spatiotemporal process is that we do not need to rely on the normality assumption for the observation process and thus nonlinear and/or non-Gaussian models could be taken into account.

For the parameters (β, α) we assume a normal prior, although the methodology described here is valid for truncated normal or improper uniform priors as well, and the variance coefficient σ^2 is assumed to be distributed according to an inverse-gamma conjugate prior, i.e.

$$\begin{aligned}
\beta|\sigma^2 &\sim N_m(Q_0^{-1}b_0, \sigma^2 Q_0^{-1}), \\
\alpha|\sigma^2 &\sim N(s_0^{-1}a_0, \sigma^2 s_0^{-1}), \\
\sigma^2 &\sim IG\left(\frac{c_0}{2}, \frac{r_0}{2}\right),
\end{aligned} \tag{3}$$

for suitable hyperparameters b_0, Q_0, a_0, s_0, c_0, and r_0. To make the priors reasonably uninformative it is common to set $Q_0 = q_0 I$, i.e. a diagonal with all diagonal elements equal to q_0, and assign small values to q_0, s_0, c_0 and r_0. If α is a matrix, as was discussed earlier, then matrix-normal prior is used instead.

2.2 Bayesian inference

We assume that observations from n fixed spatial locations are available at times $1, 2, \ldots$, but we do not assume that all n locations are observed every time. We denote by $\mathbf{y}_{1:t} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t)$ the observations up to time t where each $\mathbf{y}_s$ is an n-dimensional vector, possibly containing missing values. We denote by $\mathbf{x}_{0:t} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t)$ the value of the spatiotemporal process at the same n locations up to time t, and $\theta = (\alpha, \beta, \sigma^2)$ denotes the temporal parameters. The problem is to estimate the parameters θ and ϕ, as well as the latent process at time t, $\mathbf{x}_t$, from the available data $\mathbf{y}_{1:t}$. The problem of estimating the spatiotemporal process at locations different from the sampled locations becomes trivial after θ, ϕ, and $\mathbf{x}_{0:t}$ are estimated.

For given ϕ, we sample from the posterior distribution $p(\mathbf{x}_{0:t}, \theta | \mathbf{y}_{1:t}, \phi)$ by combining the particle filter resampling method with a skewed proposal and the sufficient statistics method of Storvik (2002) and Fearnhead (2002). The setup of Section 2.1 allows sampling from the full conditional distribution of the parameters θ via Gibbs sampling.

More precisely, the full conditional distribution of the parameter β is

$$p(\beta | \mathbf{x}_{0:t}, \mathbf{y}_{1:t}, \alpha, \sigma^2, \phi) \propto p(\beta) \prod_{s=1}^{t} p(\mathbf{x}_s | \mathbf{x}_{s-1}, \alpha, \beta, \sigma^2, \phi),$$

which is easily shown to be normal $\beta | (\mathbf{x}_{0:t}, \mathbf{y}_{1:t}, \alpha, \sigma^2, \phi) \sim N_m(Q_t^{-1} b_t, \sigma^2 Q_t^{-1})$, where

$$\begin{aligned}
Q_t &= Q_0 + G_0' R^{-1} G_0 + \sum_{s=1}^{t} (G_s - \alpha G_{s-1})' R^{-1} (G_s - \alpha G_{s-1}), \\
b_t &= b_0 + G_0' R^{-1} \mathbf{x}_0 + \sum_{s=1}^{t} (G_s - \alpha G_{s-1})' R^{-1} (\mathbf{x}_s - \alpha \mathbf{x}_{s-1}),
\end{aligned} \tag{4}$$

and similarly for α and σ^2. Note that the full conditional distributions of $\theta = (\alpha, \beta, \sigma^2)$ depend on some sufficient quantities, $u_t = u_t(\mathbf{x}_{0:t}, \phi)$, which are updated recursively. (We use the term "sufficient quantities" instead of "sufficient statistics" because they depend on the unknown parameter ϕ.) For example, to update β, from (4), we need to keep a record of the sums $\sum G_s' R^{-1} G_s$, $\sum G_s' R^{-1} G_{s-1}$, $\sum G_s' R^{-1} \mathbf{x}_s$, $\sum G_s' R^{-1} \mathbf{x}_{s-1}$, and $\sum G_{s-1}' R^{-1} \mathbf{x}_s$ where the summation is over $s = 1, \ldots, t$. Having stored the sufficient quantities $u_{t-1}(\mathbf{x}_{0:t-1}, \phi)$ at time $t - 1$, we update them by adding the corresponding terms at time t, i.e. $u_t(\mathbf{x}_{0:t}, \phi) = \mathcal{U}(u_{t-1}(\mathbf{x}_{0:t-1}, \phi), \mathbf{x}_t, \phi)$.

To sample $(\mathbf{x}_t, \theta)$ we take marginal samples from $p(\mathbf{x}_{0:t}, \theta | \mathbf{y}_{1:t}, \phi)$. The key element in the process for doing so, given a sample from $p(\mathbf{x}_{0:t-1} | \mathbf{y}_{1:t-1}, \phi)$, is

$$p(\mathbf{x}_{0:t}, \theta | \mathbf{y}_{1:t}, \phi) \propto p(\mathbf{y}_t | \mathbf{x}_t) p(\theta | \mathbf{x}_{0:t-1}, \phi) p(\mathbf{x}_t | \mathbf{x}_{t-1}, \theta, \phi) p(\mathbf{x}_{0:t-1} | \mathbf{y}_{1:t-1}, \phi),$$

which suggests a separate, two-step, update for θ and $\mathbf{x}_t$. In the first step we sample θ from $p(\theta | \mathbf{x}_{0:t-1}, \phi) = p(\theta | u_{t-1}(\mathbf{x}_{0:t-1}, \phi))$. Because this distribution is not available in closed form but the full conditional distributions for each component of θ given the other components are, sampling is done by running a few Gibbs iterations for each component of θ. The final θ obtained at the end of the sequence of Gibbs iterations, together with $\mathbf{x}_{t-1}$ and $\mathbf{y}_t$ are used to sample $\mathbf{x}_t$. To that end, let $q(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{y}_t, \theta, \phi)$ be a proposal distribution which generates N particles $\tilde{\mathbf{x}}^{(1)}, \ldots, \tilde{\mathbf{x}}^{(N)}$. Each particle carries a weight proportional to

$$w^{(i)} = \frac{p(\mathbf{y}_t | \tilde{\mathbf{x}}_t^{(i)}) p(\tilde{\mathbf{x}}_t^{(i)} | \mathbf{x}_{t-1}, \theta, \phi)}{q(\tilde{\mathbf{x}}_t^{(i)} | \mathbf{x}_{t-1}, \mathbf{y}_{1:t}, \theta, \phi)}. \tag{5}$$

Algorithm 1 One step sampling for fixed ϕ at time t.

Input: $\mathbf{y}_t$;

    Sample $\mathbf{x}_{t-1} \sim p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}, \phi)$;

    Sufficient quantities $u_{t-1} = u_{t-1}(\mathbf{x}_{0:t-1}, \phi)$.

Execute:

  1: Sample θ from $p(\theta|u_{t-1})$ by running a few Gibbs iterations.

  2: Compute the proposal $q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t, \theta, \phi)$.

  3: Sample $\tilde{\mathbf{x}}_t^{(1)}, \ldots, \tilde{\mathbf{x}}_t^{(N)} \sim q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t, \theta, \phi)$.

  4: Compute the weight $w^{(i)}$ according to (5) for $i = 1, \ldots, N$.

  5: Sample index j from $\{1, \ldots, N\}$ with weights proportional to $w^{(1)}, \ldots, w^{(N)}$.

  6: Set $\mathbf{x}_t = \tilde{\mathbf{x}}_t^{(j)}$

  7: Update the sufficient quantities $u_t = \mathcal{U}(u_{t-1}, \mathbf{x}_t, \phi)$.

Return: $\mathbf{x}_t$, u_t, θ.

The sample $\mathbf{x}_t$ is chosen from among the $\tilde{\mathbf{x}}^{(1)}, \ldots, \tilde{\mathbf{x}}^{(N)}$ with weights proportional to $w^{(1)}, \ldots, w^{(N)}$. Algorithm 1 outlines the steps from this procedure.

On the other hand, implementing the same approach for the estimation of the range parameter ϕ is far from trivial and it cannot be updated using sufficient quantities as before, making it impractical for online applications. Furthermore, each update of ϕ requires the inversion of a large matrix which could be computationally expensive when the dimension of the random field, n, is large. Therefore a technique other than Gibbs sampling (or in general an MCMC framework) is required. We circumvent this problem by using a novel particle filter with an online implementation of an empirical Bayes method. This technique is treated next in Section 3.

2.3   A skewed-normal proposal density

A measure of the quality of the proposal distribution is the effective sample size (ESS), defined as

$$\text{ESS} = \frac{(\sum_i w_t^i)^2}{\sum_i (w_t^i)^2}.$$

It can take values between 1 and N and is used for assessing the loss of variance in the importance weights (Robert and Casella, 2010, Section 4.4). A value close to N would mean that the samples are nearly equally weighted and there is diversity among the samples so no samples are lost. A value close to 1 would indicate that all but one sample will have weight close to 0 and lead to the well-known problem of sample degeneracy.

Importance sampling methods allow flexibility in the choice of the proposal distribution q but as Doucet et al. (2000) point out the optimal proposal distribution in the sense that it minimizes the variance of the importance weights is

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{1:t}, \theta, \phi) = p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t, \theta, \phi). \tag{6}$$

Although not helpful by itself, equation (6) is still useful since it can serve as a basis for deriving suboptimal proposal distributions. One way of doing this approximates (6) by a multivariate Gaussian distribution as was done by Doucet et al. (2000) in the *univariate* case. When the state process is multivariate it is imperative to use a good proposal density as this increases the effective sample size. In view of this, we introduce next a novel importance density. First, Lemma 1 shows that the optimal proposal distribution is skewed when the observation process has a skewed

distribution. This suggests using a proposal distribution which is skewed and motivates our use of the skewed-normal distribution for this purpose.

Lemma 1. *Consider the stochastic dynamics in equation* (1) *and the observation process given in equation* (2) *which corresponds to data from a general exponential family. Then the optimal proposal distribution given in equation* (6) *is skewed when the likelihood of $y_{i,t}|x_{i,t}$ is skewed.*

Proof. See Appendix. ∎

Consider first a Gaussian approximation to (6). Note that the optimal proposal is $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t, \theta, \phi) \propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta, \phi)$ and let

$$f(\mathbf{x}_t) = -\log\{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta, \phi)\}. \tag{7}$$

Next define

$$\hat{\mathbf{x}}_t = \operatorname*{argmin}_{\mathbf{x}_t} f(\mathbf{x}_t), \qquad \hat{H}_t = \nabla\nabla' f(\hat{\mathbf{x}}_t).$$

The Gaussian proposal is constructed by setting the mean equal to $\hat{\mathbf{x}}_t$ and the variance to $\hat{H}_t^{-1}$.

To capture the skewness of the distribution we use a skewed-normal copula correction to the Gaussian proposal. The probability density function (pdf) of the univariate skewed-normal distribution is (Azzalini and Capitanio, 1999)

$$\frac{2}{\omega} \psi\left(\frac{z - \xi}{\omega}\right) \Psi\left(a\frac{z - \xi}{\omega}\right)$$

where $\psi(\cdot)$ and $\Psi(\cdot)$ denote the pdf and cumulative distribution function (cdf) of the standard normal distribution respectively. The parameters ξ, $\omega > 0$, and a correspond to the location, scale, and skewness parameter respectively.

To derive the skewed-normal corrections we expand the marginals of (7) to third order terms and match the first three moments to the skewed-normal distribution. For details see Appendix B of Rue et al. (2009). Let $\tilde{\mathbf{x}}_t$ be a sample from the Gaussian approximation to (6). The idea is to transform $\tilde{\mathbf{x}}_t$ marginally using the skewed-normal correction. Ferkingstad and Rue (2015) propose two copula corrections which we also use here: a mean-only skewness correction where the proposal distribution remains Gaussian but the mean is corrected using the skewed-normal approximations to the marginals; and a mean-plus-skewness correction where the particles are sampled from the Gaussian approximation and then are marginally transformed using the skewed-normal approximation. In our simulation study (see Supplementary Materials, Section 1) we find that the two skewed-normal proposals have similar ESS but the Gaussian proposal has significantly lower ESS. We therefore recommend the mean-only corrected proposal because it is simpler than the mean-plus-skewness correction, while it also has better ESS than the Gaussian proposal.

3   A methodology for online estimation and prediction

In this section we present an empirical Bayes approach for the estimation of the range parameter ϕ. Unlike the parameter $\theta = (\alpha, \beta, \sigma^2)$, it is not possible to include ϕ as an extra step in the Gibbs algorithm without sacrificing the online feature of the method since the sufficient quantities for the update of θ depend on ϕ and consequently they must be recomputed from time one at every update of ϕ. Instead we adopt a novel empirical Bayes method in order to estimate ϕ in a way that is similar in spirit to Doss (2010). Another argument in favor of the empirical Bayes approach

instead of a full Bayesian approach is that it is unclear how a suitable prior for ϕ should be chosen. Berger et al. (2001) discuss some objective priors in the case of Gaussian responses. However for non-Gaussian data these priors, and indeed any improper prior, result in an improper posterior for ϕ (Christensen et al., 2000). When it comes to online inference, it is unclear how a fully Bayesian approach would be implemented. If a Monte-Carlo algorithm is used, the sufficient quantities will be computed for those ϕ values in the Monte-Carlo sample only. This restricts the ϕ values at subsequent times to only those which were sampled at all previous time points, something that is undesirable. The goodness-of-fit of the empirical Bayes method for estimating the range parameter has been demonstrated in the case of the spatial-only model by Roy et al. (2016). Extending it to an online version requires careful treatment of the sufficient quantities needed to compute the Bayes factors. Theorem 1 presents the main result of this section. The approach discussed below may be also viewed as equivalent to the maximum likelihood estimation for ϕ after integrating out the parameter θ.

We consider first the marginal density $p(\mathbf{y}_{1:t}|\phi)$ and define the estimator for ϕ at time t given data $\mathbf{y}_{1:t}$ by

$$\hat{\phi}_t = \operatorname*{argmax}_{\phi} p(\mathbf{y}_{1:t}|\phi), \tag{8}$$

where

$$p(\mathbf{y}_{1:t}|\phi) = \int p(\mathbf{y}_{1:t}, \mathbf{x}_{0:t}, \theta|\phi)d(\mathbf{x}_{0:t}, \theta). \tag{9}$$

In general, the integral in (9) has no closed form and thus a numerical approximation must be employed. Define the sequential Bayes factor between ϕ and $\tilde{\phi}$ with respect to the data $\mathbf{y}_{1:t}$ by

$$B_{1:t}(\phi; \tilde{\phi}) = \frac{p(\mathbf{y}_{1:t}|\phi)}{p(\mathbf{y}_{1:t}|\tilde{\phi})}.$$

Note the dependence of the Bayes factor on the whole data sequence $\mathbf{y}_{1:t}$. Then, for a fixed parameter $\tilde{\phi}$, (8) is equivalent to

$$\hat{\phi}_t = \operatorname*{argmax}_{\phi} B_{1:t}(\phi; \tilde{\phi}).$$

Furthermore, the sequential Bayes factor, $B_{1:t}(\phi; \tilde{\phi})$ in a filtering framework is computed as follows:

$$\begin{aligned} B_{1:t}(\phi; \tilde{\phi}) &= \int \frac{p(\mathbf{y}_{1:t}, \mathbf{x}_{0:t}, \theta|\phi)}{p(\mathbf{y}_{1:t}, \mathbf{x}_{0:t}, \theta|\tilde{\phi})} p(\mathbf{x}_{0:t}, \theta|\mathbf{y}_{1:t}, \tilde{\phi})d(\mathbf{x}_{0:t}, \theta) \\ &= \int \frac{p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t})p(\mathbf{x}_{0:t}|\theta, \phi)p(\theta)}{p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t})p(\mathbf{x}_{0:t}|\theta, \tilde{\phi})p(\theta)} p(\mathbf{x}_{0:t}, \theta|\mathbf{y}_{1:t}, \tilde{\phi})d(\mathbf{x}_{0:t}, \theta) \\ &= \int \frac{p(\mathbf{x}_{0:t}|\theta, \phi)}{p(\mathbf{x}_{0:t}|\theta, \tilde{\phi})} p(\mathbf{x}_{0:t}, \theta|\mathbf{y}_{1:t}, \tilde{\phi})d(\mathbf{x}_{0:t}, \theta). \end{aligned} \tag{10}$$

A naive approach for estimating ϕ relying on equation (10) would be to obtain a large sample for $(\mathbf{x}_{0:t}, \theta)$ from $p(\mathbf{x}_{0:t}, \theta|\mathbf{y}_{1:t}, \tilde{\phi})$ using Algorithm 1, and to approximate (10) by Monte-Carlo integration; call the result $\hat{B}_{1:t}(\phi; \tilde{\phi})$. Then an estimate would be obtained by maximizing $\hat{B}_{1:t}(\phi; \tilde{\phi})$ over ϕ.

Remark 1. *There are several issues concerning the above naive approach that need to be addressed:*

1. *Unless $\hat{\phi}_t$ and $\tilde{\phi}$ are close to one another, the Monte-Carlo approximation will have a large error and in this case the estimate may not be accurate no matter how large the Monte-Carlo sample is.*

8

2. *In order to compute $p(\mathbf{x}_{0:t}|\theta,\phi)$ for any ϕ at time-point t, we need $p(\mathbf{x}_{0:t-1}|\theta,\phi)$ for the same ϕ at time $t-1$ something that we could not anticipate prior to time t.*

3. *After obtaining $\hat{\phi}_t$ we need to run Algorithm 1 once more in order to update $\mathbf{x}_t$ and θ conditional on $\hat{\phi}_t$. However the algorithm requires samples from $\mathbf{x}_{0:t-1}|\mathbf{y}_{1:t-1},\hat{\phi}_t$ which is unknown at time $t-1$.*

Bypassing the issues raised in Remark 1, our strategy for resolving issue 1 is to replace the importance density in (10) by a mixture over a set of ϕ values instead of a single fixed $\tilde{\phi}$. This is demonstrated in our simulation study (see Supplementary Materials, Section 2) where we show that the naive approach can give biased estimates for a badly chosen $\tilde{\phi}$. For 2 we only evaluate the Bayes factors over a dense grid Φ which covers reasonable values of ϕ and keeps a record of the sufficient quantities needed to evaluate $p(\mathbf{x}_{0:t}|\theta,\phi)$ at the next time. To resolve 3 we resample the available samples from the mixture with appropriate weights. In the remainder of this section we expand on these ideas. Algorithm 2 puts them together.

Consider a set $\Phi_K = \{\phi_1,\ldots,\phi_K\}$ such that $\tilde{\phi} \in \Phi_K$, is sufficiently spread-out over a range of interesting values of ϕ. The meaning of "interesting values of ϕ" is well defined in our context: the range parameter is a scaling factor of the spatial distances within the domain of interest which define a possible range for ϕ. Suppose $(\mathbf{x}_{0:t}^{(l,k)}, \theta^{(l,k)})$, $k = 1,\ldots,K$, $l = 1,\ldots,L_k$ are samples from $p(\mathbf{x}_{0:t},\theta|\mathbf{y}_{1:t},\phi_k)$. The augmented sample can be seen as drawn from the mixture distribution

$$p_{\mathrm{mix}}(\mathbf{x}_{0:t},\theta|\mathbf{y}_{1:t},\Phi_K,\Lambda_K) = \sum_{k=1}^{K} \lambda_k p(\mathbf{x}_{0:t},\theta|\mathbf{y}_{1:t},\phi_k), \tag{11}$$

where $\lambda_k = L_k/(\sum L_{k'})$ and $\Lambda_K = \{\lambda_1,\ldots,\lambda_K\}$. Let $b_t^k = B_{1:t}(\phi_k;\tilde{\phi})$. Then b_t^k can be estimated by maximizing the so-called reverse logistic log-likelihood (Geyer, 1994)

$$\ell(\mathbf{b}_t) = \sum_{k=1}^{K} \sum_{l=1}^{L_k} \log \frac{\lambda_k p(\mathbf{x}_{0:t}^{(k,l)}, \theta^{(k,l)}|\mathbf{y}_{1:t},\phi_k)}{p_{\mathrm{mix}}(\mathbf{x}_{0:t}^{(k,l)}, \theta^{(k,l)}|\mathbf{y}_{1:t},\Phi_K,\Lambda_K)}.$$

Furthermore, let $\hat{b}_t^k$ denote the estimate for b_t^k. Then, the following sum

$$\hat{B}_{1:t}(\phi;\tilde{\phi}) = \sum_{k=1}^{K} \sum_{l=1}^{L_k} \frac{p(\mathbf{x}_{0:t}^{(k,l)}|\theta^{(k,l)},\phi)}{\sum_{k'} L_{k'}/\hat{b}_t^{k'} p(\mathbf{x}_{0:t}^{(k,l)}|\theta^{(k,l)},\phi_{k'})}, \tag{12}$$

estimates $B_{1:t}(\phi;\tilde{\phi})$. The key Theorem 1 summarizes this property.

Theorem 1. *Consider a coarse grid $\Phi_K = \{\phi_1,\phi_2,\ldots,\phi_K\}$, where the grid points, ϕ_k, $k = 1,\ldots,K$, are spaced across the parameter space for ϕ. Suppose that for $k = 1,\ldots,K$, we draw samples $(\mathbf{x}_{1:t}^{(k,l)}, \theta^{(k,l)})$, $l = 1,\ldots,L_k$ from the distribution $p(\mathbf{x}_{0:t},\theta|\phi_k,\mathbf{y}_{1:t})$ for $\phi_k \in \Phi_K$. Then, for an arbitrarily fixed pair $(\phi,\tilde{\phi})$, the estimate,*

$$\hat{B}_{1:t}(\phi;\tilde{\phi}) \xrightarrow{a.s.} B_{1:t}(\phi;\tilde{\phi}), \quad L_k \to \infty, \tag{13}$$

where $\hat{B}_{1:t}(\phi;\tilde{\phi})$ is given by equation (12).

Proof. See Appendix.   ∎

The likelihood in the numerator and denominator of (12) must be computed for the whole history of samples $\mathbf{x}_{0:t}^{(k,l)}$ for the new $\theta^{(k,l)}$ and for different values of ϕ. This is not as straightforward as a product of the prior $p(\mathbf{x}_{0:t-1}|\theta,\phi)$ times the transition $p(\mathbf{x}_t|\mathbf{x}_{t-1},\theta,\phi)$ since in online implementations we cannot anticipate the value of θ and ϕ at time $t-1$. If ϕ is held fixed, then $p(\mathbf{x}_{0:t}|\theta,\phi)$ can be written in terms of sufficient quantities as before which do not depend on θ but do depend on ϕ.

To this end, let $z_{t-1}^{(\phi,k,l)}$ denote the aforementioned sufficient quantities at time $t-1$ for the (k,l)th sample. These are updated at time t by

$$z_t^{(\phi,k,l)} = \mathcal{Z}\left(z_{t-1}^{(\phi,k,l)}, \mathbf{x}_t^{(k,l)}, \phi\right),$$

and the joint likelihood is expressed as a function of $z_t^{(\phi,k,l)}$ and $\theta^{(k,l)}$, i.e.

$$p(\mathbf{x}_{0:t}^{(k,l)}|\theta^{(k,l)},\phi) = \mathcal{P}\left(z_t^{(\phi,k,l)}, \theta^{(k,l)}\right).$$

In practice, we do not consider the entire parameter space for ϕ but a fine discretization of it. More precisely, we augment the coarse grid, Φ_K, with the finer grid, say Φ, i.e. $\Phi_K \subset \Phi$, and compute the sufficient quantities $z_t^{(\phi,k,l)}$, and consequently $\hat{B}_{1:t}(\phi;\tilde{\phi})$, only for those $\phi \in \Phi$. Then, we estimate ϕ by

$$\hat{\phi}_t = \underset{\phi\in\Phi}{\operatorname{argmax}}\, \hat{B}(\phi;\tilde{\phi}).$$

The drawback is the loss of precision in estimation but the benefit is that the required computing memory remains fixed. From experience, we find that a small bias in the value of ϕ does not affect prediction or the estimation of the other parameters. In particular in the simulation study presented in Section 2 of the Supplementary Materials we find that the parameters α and β are immune to the possible bias in $\hat{\phi}_t$.

The empirical Bayes approach proceeds with the update of the estimate for $(\mathbf{x}_t,\theta)$ using the estimate $\hat{\phi}_t$. These estimates are obtained as a weighted sum of the existing samples $(\mathbf{x}_{0:t}^{(k,l)},\theta^{(k,l)})$, $k=1,\ldots,K$, $l=1,\ldots,L_k$. The distribution of the existing samples is the mixture distribution (11). These samples can be scaled with reference to the distribution conditioned on $\phi = \hat{\phi}_t$ using the following importance weights

$$
\begin{aligned}
v_t^{(k,l)} &= \frac{p(\mathbf{x}_{0:t}^{(k,l)}, \theta^{(k,l)}|\mathbf{y}_{1:t}, \hat{\phi}_t)}{p_{\text{mix}}(\mathbf{x}_{0:t}^{(k,l)}, \theta^{(k,l)}|\mathbf{y}_{1:t}, \Phi_K, \Lambda_K)} = \frac{p(\mathbf{x}_{0:t}^{(k,l)}, \theta^{(k,l)}|\mathbf{y}_{1:t}, \hat{\phi}_t)}{\sum_{k'} \lambda_{k'} p(\mathbf{x}_{0:t}^{(k,l)}, \theta^{(k,l)}|\mathbf{y}_{1:t}, \phi_{k'})} \\
&= \frac{p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t}^{(k,l)})p(\mathbf{x}_{0:t}^{(k,l)}|\theta^{(k,l)}, \hat{\phi}_t)p(\theta^{(k,l)})/p(\mathbf{y}_{1:t}|\hat{\phi}_t)}{\sum_{k'} \lambda_{k'} p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t}^{(k,l)})p(\mathbf{x}_{0:t}^{(k,l)}|\theta^{(k,l)}, \phi_{k'})p(\theta^{(k,l)})/p(\mathbf{y}_{1:t}|\phi_{k'})} \\
&= \frac{p(\mathbf{x}_{0:t}^{(k,l)}|\theta^{(k,l)}, \hat{\phi}_t)/B_{1:t}(\hat{\phi}_t;\tilde{\phi})}{\sum_{k'} \lambda_{k'} p(\mathbf{x}_{0:t}^{(k,l)}|\theta^{(k,l)}, \phi_{k'})/B_{1:t}(\phi_{k'};\tilde{\phi})}.
\end{aligned}
\tag{14}
$$

In practice $B_{1:t}(\hat{\phi}_t;\tilde{\phi})$ and $B_{1:t}(\phi_{k'};\tilde{\phi})$ in equation (14) are replaced by their estimates which are already available. Then, we obtain the estimates for the latent state process and the remaining parameters by

$$\hat{\mathbf{x}}_t = \sum_{k=1}^{K}\sum_{l=1}^{L_k} \bar{v}_t^{(k,l)}\mathbf{x}_t^{(k,l)}, \quad \hat{\theta}_t = \sum_{k=1}^{K}\sum_{l=1}^{L_k} \bar{v}_t^{(k,l)}\theta^{(k,l)},$$

Algorithm 2 Main estimation and prediction algorithm at time t.

Input: $\mathbf{y}_t$;

Samples $\mathbf{x}_{t-1}^{(k,1:L_k)} \sim p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}, \phi_k)$, $k = 1, \ldots, K$;

Sufficient quantities $u_{t-1}^{(k,1:L_k)}$, $k = 1, \ldots, K$;

Sufficient quantities $z_{t-1}^{(\phi,k,1:L_k)}$, $\phi \in \Phi$, $k = 1, \ldots, K$.

Execute:

1: **for** $k \in \{1, \ldots, K\}$ **do concurrently**
2:    **for** $l = 1, \ldots, L_k$ **do concurrently**
3:       Sample l' uniformly in $\{1, \ldots, L_k\}$.
4:       Call Algorithm 1 with input $\mathbf{y}_t$, $\mathbf{x}_{t-1}^{(k,l')}$, $u_{t-1}^{(k,l')}$, and output $\theta^{(k,l)}$, $\mathbf{x}_t^{(k,l)}$, $u_t^{(k,l)}$.
5:       **for** $\phi \in \Phi$ **do concurrently**
6:          Update the sufficient quantities $z_t^{(\phi,k,l)} = \mathcal{Z}\left(z_{t-1}^{(\phi,k,l)}, \mathbf{x}_t^{(k,l)}, \phi \right)$.
7:          Compute $p(\mathbf{x}_{0:t}^{(k,l)}|\theta^{(k,l)}, \phi) = \mathcal{P}\left(z_t^{(\phi,k,l)}, \theta^{(k,l)} \right)$.
8:       **end for** ϕ
9:    **end for** l
10: **end for** k
11: Call the reverse logistic regression algorithm with input $\{p(\mathbf{x}_{0:t}^{(k,1:L_k)}|\theta^{(k,1:L_k)}, \phi) : \phi \in \Phi, k = 1, \ldots, K\}$ and output $\hat{\mathbf{b}}_t$.
12: Compute $\hat{B}_{1:t}(\phi; \tilde{\phi})$, $\phi \in \Phi$ using (12).
13: Set $\hat{\phi}_t = \operatorname{argmax}_{\phi \in \Phi} \hat{B}_{1:t}(\phi; \tilde{\phi})$.
14: Compute importance weights $v_t^{(k,l)}$ according to equation (14) and normalize them to get $\bar{v}_t^{(k,l)}$.
15: Set $\hat{\mathbf{x}}_t = \sum_{k=1}^K \sum_{l=1}^{L_k} \bar{v}_t^{(k,l)} \mathbf{x}_t^{(k,l)}$, $\hat{\theta}_t = \sum_{k=1}^K \sum_{l=1}^{L_k} \bar{v}_t^{(k,l)} \theta^{(k,l)}$.

Return: $\hat{\mathbf{x}}_t$, $\hat{\theta}_t$, $\hat{\phi}_t$,

$\mathbf{x}_t^{(k,1:L_k)}$, $\theta^{(k,1:L_k)}$, $u_t^{(k,1:L_k)}$, $k = 1, \ldots, K$,

$z_t^{(\phi,k,1:L_k)}$, $\phi \in \Phi$, $k = 1, \ldots, K$.

---

where $\bar{v}_t^{(k,l)}$ is the normalized version of (14). This is the final step at time t. The main algorithm of this paper which shows how to combine Algorithm 1 with the online empirical Bayes for the estimation of ϕ is displayed in Algorithm 2.

We evaluated the performance of our method in a simulation study which we include in the Supplementary Materials with this article. Our study shows that the obtained parameter estimates are unbiased and consistent. As more data are assimilated, the credible intervals obtained from the Monte-Carlo samples become narrower as expected with rate in the order of the square root of the number of the elapsed time. The posterior distributions obtained by our method, conditioned on the data $\mathbf{y}_{1:t}$, are compared against those derived from a typical offline, fully Bayesian, MCMC method. Indeed, the two distributions match which indicates the Monte-Carlo samples are taken from the correct distribution. Overall, this study verifies our theoretical conclusions and justifies the use of our method for online inference.

4   Example: Spatiotemporal monitoring of the Cs-137 isotope

The Fukushima Daiichi nuclear disaster was a catastrophic failure at the Fukushima Nuclear Power Plant on 11 March 2011, resulting in a meltdown of three of the plant's six nuclear reactors. The failure occurred when the plant was hit by the tsunami following an earthquake. The Japanese

authorities started to collect data about the radioactive material released from the power station which were reported to the International Atomic Energy Agency (IAEA). At an early stage of the accident, online methods were needed to incorporate new measurements in real-time. The data analyzed in this paper consist of daily measurements of radioactive decay for the Caesium-137 (Cs-137) isotope found on leaves collected between 16 March 2011 and 26 December 2011. The measurements were collected from different locations and the number of nuclear decays of the isotope in one second were counted. We refer the reader to the IAEA relevant website `https://iec.iaea.org/fmd` for more information and access to the datasets.

The samples were taken at $n = 17$ distinct locations across $T = 146$ days. However some locations were sampled more than once on the same day so the total of all the measurements were used for that day. An intercept term, a time trend, and the distance from the power plant were used as covariates. Our aim is to estimate the parameters $(\alpha, \beta, \sigma^2, \phi)$ as well as predict the spatiotemporal field $\mathbf{x}_t$ at time t from observations $\mathbf{y}_{1:t}$. In other words the hidden spatiotemporal process is given by

$$\mathbf{x}_t = \beta_0 + \beta_1 g + \beta_2 t + \eta_t,$$
$$\eta_t = \alpha \eta_{t-1} + \epsilon_t, \tag{15}$$

where g is the distance from the station, $\epsilon_t \sim \mathcal{N}(0, \sigma^2 R(\phi))$, and $\mathrm{Corr}(x_{i,t}, x_{j,t}) = e^{-d_{ij}/\phi}$. Moreover, the ith collected observation is conditionally distributed according to,

$$y_{i,t}|x_{i,t} \sim \mathrm{Poisson}(\tau_{i,t} e^{x_{i,t}}),$$

where $\tau_{i,t}$ corresponds to the number of times that location i was sampled at day t. These locations are shown in Figure 1.

The priors for $(\alpha, \beta, \sigma^2)$ were used as in Section 2 with the following parameters: $a_0 = 0$, $s_0 = 0.1$, $b_0 = 0$, $q_0 = 0.01$, $d_0 = 0.1$, $e_0 = 0.1$. The fine grid Φ for estimating ϕ consists of 51 equally spaced points in $[0, 0.1]$ and the coarse grid Φ_K consists of 7 equally spaced points in $[0.002, 0.098]$. Algorithm 2 was used for estimation and prediction with particle size $N = 1000$, MCMC size $L = 500$ and Gibbs burn-in $B = 100$.

Figure 2 shows the evolution of the parameter estimates in time. There is an apparent "jump" in the parameter estimates for β and ϕ at around time $t = 70$ after which the estimates become stable. A closer examination of the data reveals that this may be due to a lower trend after time 50 and to reduced sampling after time 40.

Figure 3 shows the prediction at 2659 locations around the sampling area for selected times. To sample from the unmonitored locations we simulate from its conditional distribution given the samples at the monitored locations and the parameters, $p(\mathbf{x}_t^*|\mathbf{x}_{0:t}, \theta, \phi)$, where $\mathbf{x}_t^*$ is the value of the state process at the prediction locations. Figure 4 shows the standard deviation of our predictions. From the plots we can identify some radiation hot-spots and an apparent decrease of radiation over time.
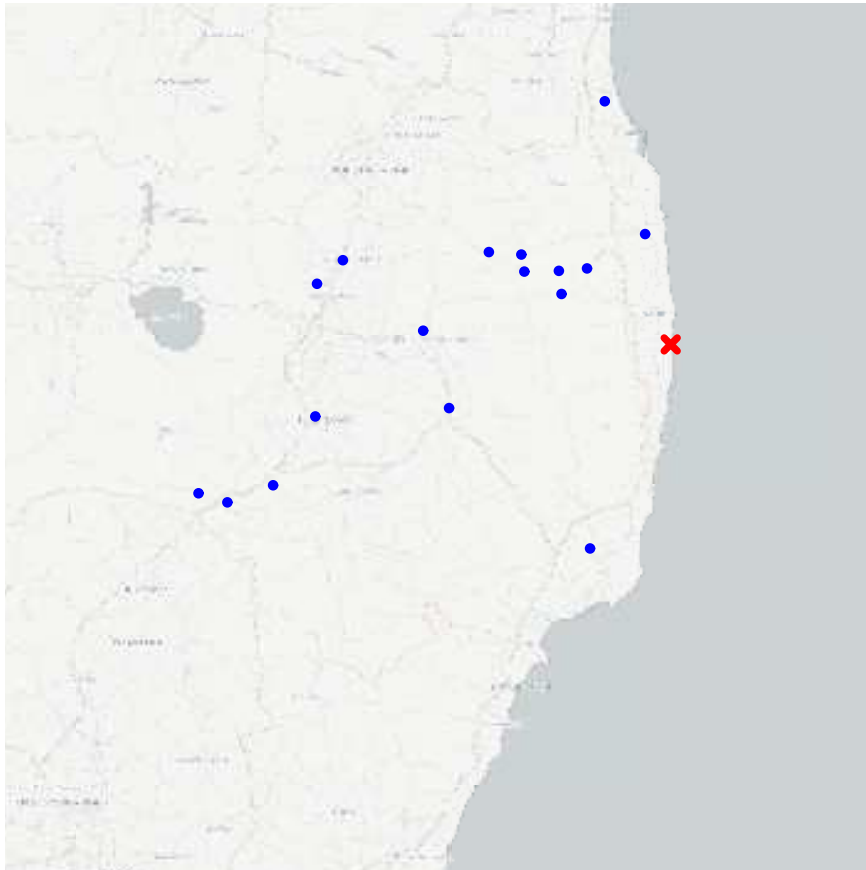
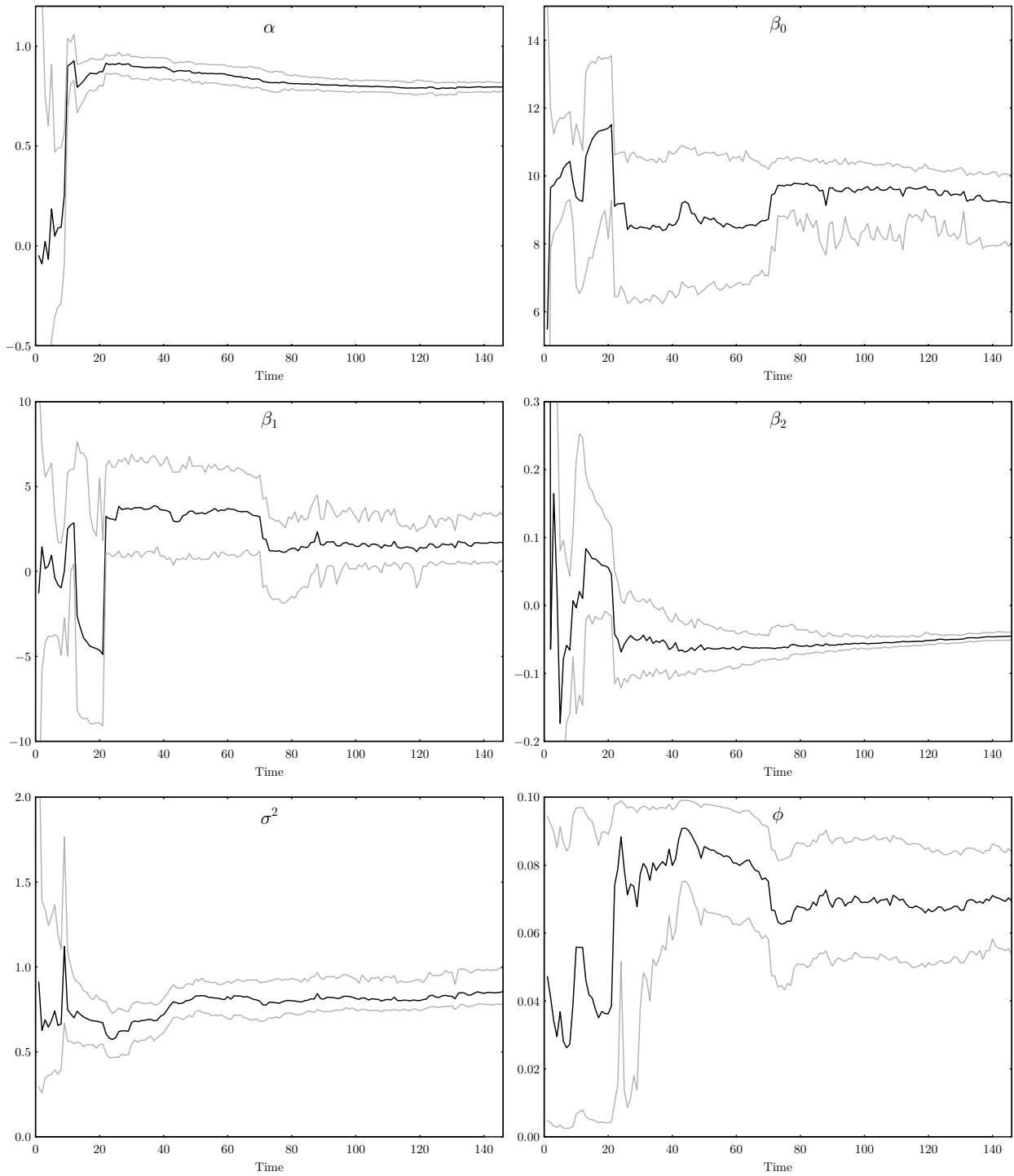Figure 1: Sampled locations of the Cs-137 isotope, shown by a ●. The location of the power plant is shown by a ×.

Figure 2: Estimates and 90% credible intervals for the parameters of the Fukushima power plant example.
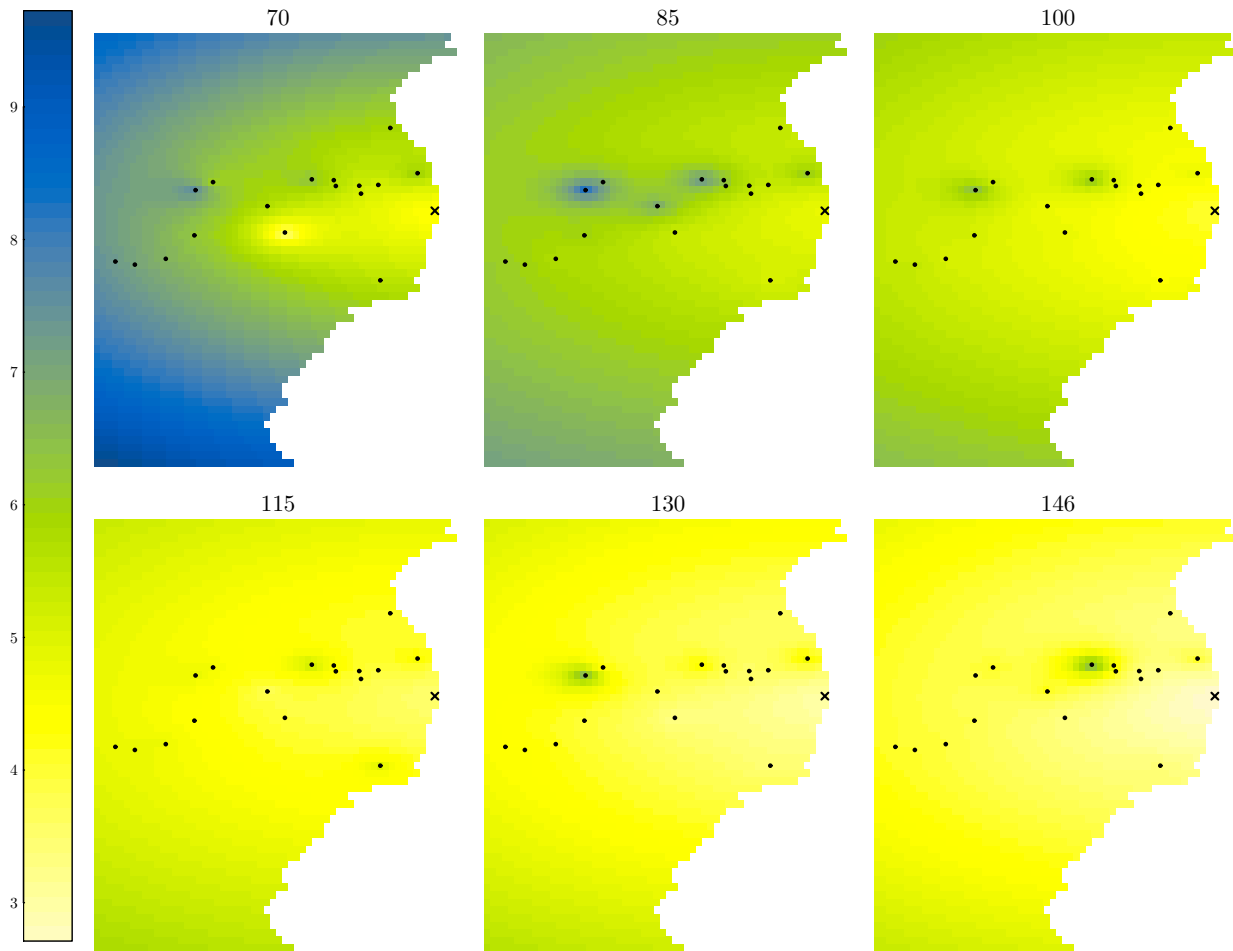
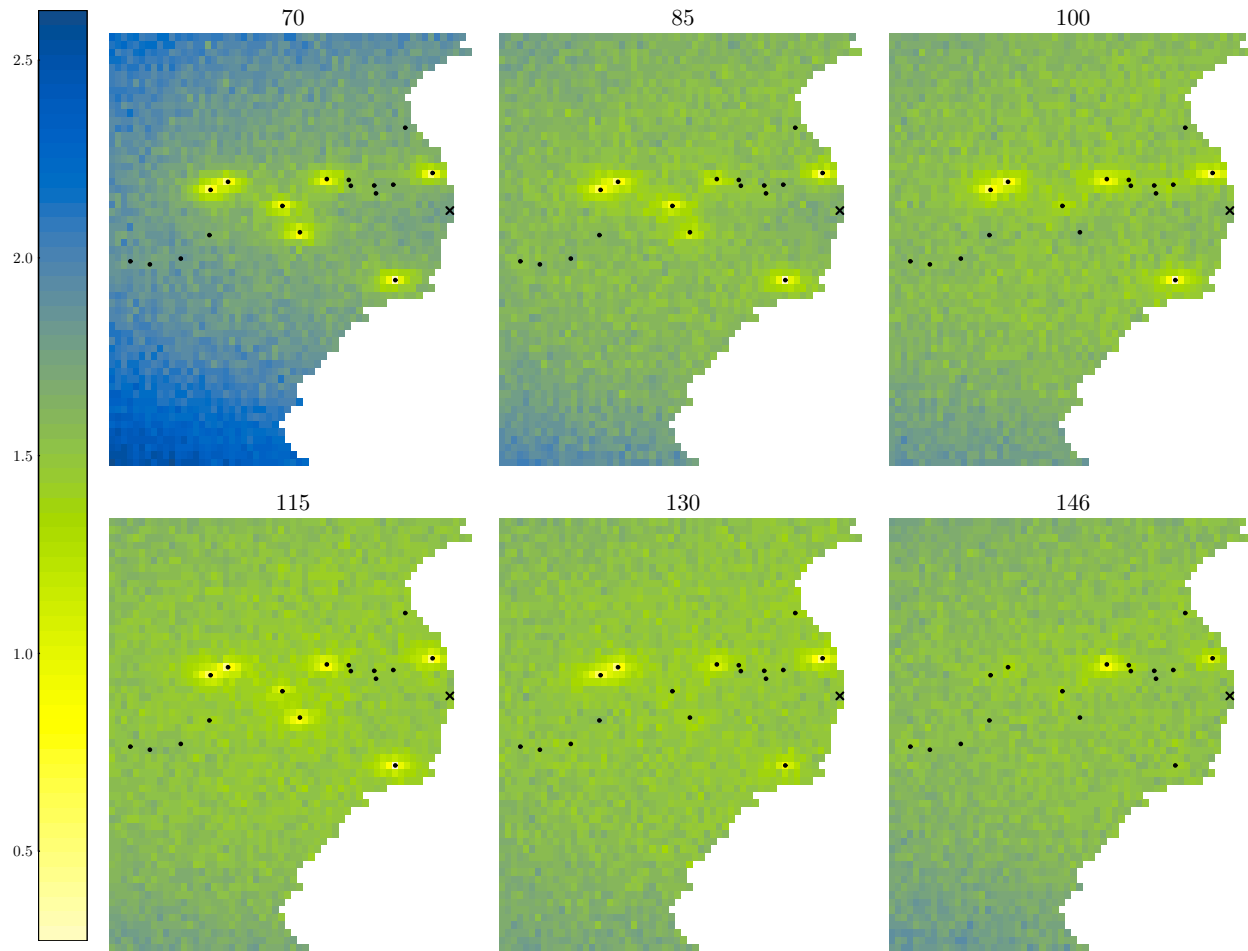Figure 3: Predictions of the state process at different times for the Fukushima power plant example.

Figure 4: Prediction standard deviation of the state process at different times for the Fukushima power plant example.

5 Summary and discussion

In this paper we propose a method for online estimation and prediction of dynamic spatiotemporal processes. We consider a latent Gaussian autoregressive spatial process with data arriving sequentially in time with distribution from an exponential family conditional on the spatiotemporal process. Our model is expressed in terms of unknown parameters which are estimated along with the latent process within an empirical Bayes framework. We distinguish two types of parameters, the temporal parameters, which have a full conditional distribution that can be written in terms of sufficient quantities, and the spatial correlation parameters, which don't. The spatial range parameter belongs to the latter type.

An algorithm is proposed for sampling from the filtering distribution of the spatiotemporal process and the posterior distribution of those parameters whose full conditional can be written in terms of sufficient quantities. These sufficient quantities are updated when new samples are taken for a fixed value of the range parameter, and because the filtering distribution can be skewed, we show how to use the skew-normal distribution to generate good candidate samples. The advantage of using sufficient quantities is that the storage requirements do not increase in time. These sufficient quantities depend on the spatial range parameter, thus, they are computed at a fixed set of values of that parameter across different times. Estimation of the range parameter is performed by maximizing the Bayes factors over this fixed set. The Bayes factors are estimated sequentially by importance sampling using the Monte Carlo samples.

Our method was compared against a typical offline MCMC method which samples from the posterior distribution of all parameters and the spatiotemporal field. We find that the distribution of the samples from our method matches the one obtained when the offline method is used. Finally, we demonstrate the application of our method on radiation measurements from a nuclear accident which can be used to assess the radiation risk in real time and provide helpful insight about its distribution.

Although the empirical Bayes estimation was applied to a single spatial correlation parameter, the theory is more general to allow more parameters to be estimated this way, e.g. a smoothness or a nugget parameter. For an application of this approach to the isotropic spatial model see Roy et al. (2016). On the other hand, when many parameters are included, the sampling and evaluation grids must be chosen carefully as a larger grid takes longer to compute.

Another extension of our model, is the use of a spatially-varying temporal autocorrelation parameter. Although this model allows us to capture the influence across spatial components, it becomes challenging to fit as there are more parameters to estimate.

Our model for the latent process assumes linear transition in time. Although not considered in this paper, it would be possible to apply the methodology to non-linear models using local linear approximations.

A potential research avenue is the application of this methodology to the dynamic spatiotemporal design problem, see e.g. Wikle and Royle (1999) incorporating parameter uncertainty in the design as well. Many interesting applications can be found in the point-process framework and it would be interesting to see how the suggested methodology performs in this case. Finally, the ideas of this paper can be applied to other models beyond the spatial framework.

6 Appendix

6.1 Proof of Lemma 1

We examine the limits for $i = 1, \ldots, n$, of the ratio

$$\lim_{u \to \infty} \frac{p(\mathbf{x}_t = \mu_t + u\mathbf{e}_i | \mathbf{x}_{t-1}, \mathbf{y}_t, \theta, \phi)}{p(\mathbf{x}_t = \mu_t - u\mathbf{e}_i | \mathbf{x}_{t-1}, \mathbf{y}_t, \theta, \phi)},$$

where $\mu_t = G_t \beta + \alpha(\mathbf{x}_{t-1} - G_{t-1}\beta)$ and $\mathbf{e}_i$ is a vector whose ith component is 1 and all other components are 0. If the limit is 0 or ∞, then the distribution is left or right skewed respectively.

Then, by the symmetry of the normal distribution around its mean,

$$\lim_{u \to \infty} \frac{p(\mathbf{x}_t = \mu_t + u\mathbf{e}_i | \mathbf{x}_{t-1}, \mathbf{y}_t, \theta, \phi)}{p(\mathbf{x}_t = \mu_t - u\mathbf{e}_i | \mathbf{x}_{t-1}, \mathbf{y}_t, \theta, \phi)} = \lim_{u \to \infty} \frac{p(\mathbf{y}_t | \mathbf{x}_t = \mu_t + u\mathbf{e}_i)p(\mathbf{x}_t = \mu_t + u\mathbf{e}_i | \mathbf{x}_{t-1}, \theta, \phi)}{p(\mathbf{y}_t | \mathbf{x}_t = \mu_t - u\mathbf{e}_i)p(\mathbf{x}_t = \mu_t - u\mathbf{e}_i | \mathbf{x}_{t-1}, \theta, \phi)}$$

$$= \lim_{u \to \infty} \frac{p(\mathbf{y}_t | \mathbf{x}_t = \mu_t + u\mathbf{e}_i)}{p(\mathbf{y}_t | \mathbf{x}_t = \mu_t - u\mathbf{e}_i)}$$

$$= \lim_{u \to \infty} \frac{p(y_{i,t} | x_{i,t} = \mu_{i,t} + u)}{p(y_{i,t} | x_{i,t} = \mu_{i,t} - u)},$$

where the last limit is either 0 or ∞ since the distribution of $y_{i,t} | x_{i,t}$ is skewed.

6.2 Proof of Theorem 1

The estimate of the sequential empirical Bayes factor, $\hat{B}_{1:t}(\phi; \tilde{\phi})$ in equation (12), depends on the associated sequential empirical Bayes factors, $\mathbf{b}_t = (b_t^1, \ldots, b_t^K)$, on the coarse grid for $\phi_k \in \Phi_K$. Consequently, we need to first establish the convergence of $\hat{b}_t^k$, $k = 1, \ldots, K$.

Because $\theta^{(k,l)}$ is drawn using Gibbs sampling, and because $\mathbf{x}_t^{(k,l)}$ is sampled by importance sampling conditioned on $\theta^{(k,l)}$, the sample $(\mathbf{x}_{0:t}^{(k,l)}, \theta^{(k,l)})$, $l = 1, \ldots, L_k$ is a Harris ergodic Markov chain for each $k \in \{1, \ldots, K\}$ from the distribution $p(\mathbf{x}_{0:t}, \theta | \mathbf{y}_{1:t}, \phi_k)$.

Let $\lambda_k = L_k / \sum L_{k'}$ and $\Lambda_K = \{\lambda_1, \ldots, \lambda_K\}$. Then the concatenated sample $(x_{1:t}^{(k;l)}, \theta^{(k;l)})$, $l = 1, \ldots, L_k$, $k = 1, \ldots, K$ is a Harris ergodic Markov chain from the mixture distribution with components the $p(\mathbf{x}_{0:t}, \theta | \mathbf{y}_{1:t}, \phi_k)$ and corresponding weights λ_k. The probability that the (k,l)th sample is drawn from the kth mixture component is given by

$$f(\mathbf{x}_{0:t}^{(k,l)}, \theta^{(k,l)} | \mathbf{y}_{1:t}, \phi_k) = \frac{\lambda_k p(\mathbf{x}_{0:t}^{(k,l)}, \theta^{(k,l)} | \mathbf{y}_{1:t}, \phi_k)}{p_{\mathrm{mix}}(\mathbf{x}_{0:t}^{(k,l)}, \theta^{(k,l)} | \mathbf{y}_{1:t}, \Phi_K, \Lambda_K)},$$

where $p_{\mathrm{mix}}(\mathbf{x}_{0:t}, \theta | \mathbf{y}_{1:t}, \Phi_K, \Lambda_K)$ denotes the mixture distribution of $p(\mathbf{x}_{0:t}, \theta | \mathbf{y}_{1:t}, \phi_k)$ for $k = 1, \ldots, K$ with weights λ_k defined in (11). Define the reverse logistic log-likelihood

$$\ell(\mathbf{b}_t) = \sum_{k=1}^{K} \sum_{l=1}^{L_k} \log f(\mathbf{x}_{0:t}^{(k,l)}, \theta^{(k,l)} | \mathbf{y}_{1:t}, \phi_k). \tag{16}$$

Then using similar arguments as in Buta and Doss (2011) one may show that the maximizing argument of ℓ, i.e. $\hat{\mathbf{b}}_t = \mathrm{argmax}\, \ell(\mathbf{b}_t)$ converges a.s. to the sequential empirical Bayes factors $\mathbf{b}_t$.

Next, observe that $\hat{B}_{1:t}(\phi; \tilde{\phi})$ can be written as

$$\sum_{k=1}^{K} \frac{1}{L_k} \sum_{l=1}^{L_k} \frac{\lambda_k p(\mathbf{x}_{0:t}^{(k,l)} | \theta^{(k,l)}, \phi)}{\sum_{k'=1}^{K} \frac{\lambda_{k'}}{b_t^{k'}} p(\mathbf{x}_{0:t}^{(k,l)} | \theta^{(k,l)}, \phi_{k'})} \xrightarrow{\text{a.s.}} \sum_{k=1}^{K} \int \frac{\lambda_k p(\mathbf{x}_{0:t} | \theta, \phi)}{\sum_{k'=1}^{K} \frac{\lambda_{k'}}{b_t^{k'}} p(\mathbf{x}_{0:t} | \theta, \phi_{k'})} p(\mathbf{x}_{0:t}, \theta | \mathbf{y}_{1:t}, \phi_k) d(\mathbf{x}_{0:t}, \theta).$$

$$\tag{17}$$

The right hand side of equation (17) equals

$$B_{1:t}(\phi; \tilde{\phi}) \times \sum_{k=1}^{K} \int \frac{\lambda_k p(\mathbf{x}_{0:t}|\theta, \phi)/p(\mathbf{y}_{1:t}|\phi)}{\sum_{k'=1}^{K} \lambda_{k'} p(\mathbf{x}_{0:t}|\theta, \phi_{k'})/p(\mathbf{y}_{1:t}|\phi_{k'})} p(\mathbf{x}_{0:t}, \theta|\mathbf{y}_{1:t}, \phi_k) d(\mathbf{x}_{0:t}, \theta), \qquad (18)$$

and multiplying and dividing by $p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t})p(\theta)$, one deduces that the finite sum of equation (18) equals 1. The proof is thus complete.

References

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602.

Berger, J. O., De Oliveira, V., and Sansó, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456):1361–1374.

Buta, E. and Doss, H. (2011). Computational approaches for empirical bayes method and bayesian sensitivity analysis. *The Annals of Statistics*, 39:2658–2685.

Christensen, O. F., Møller, J., and Waagepetersen, R. (2000). Analysis of spatial data using generalized linear mixed models and Langevin-type Markov chain Monte Carlo. Technical report, Department of Mathematical Sciences, Aalborg University.

Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Wiley.

Doss, H. (2010). Estimation of large families of Bayes factors from Markov chain output. *Statistica Sinica*, 20(2):537–560.

Doucet, A., Freitas, N., and Gordon, N., editors (2001). *Sequential Monte Carlo Methods in Practice*. Information Science and Statistics. Springer, New York.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208.

Fearnhead, P. (2002). Markov chain Monte Carlo, sufficient statistics, and particle filters. *Journal of Computational and Graphical Statistics*, 11(4):848–862.

Ferkingstad, E. and Rue, H. (2015). Improving the INLA approach for approximate Bayesian inference for latent Gaussian models. *Electronic Journal of Statistics*, 9(2):2706–2731.

Geyer, C. J. (1994). Estimating normalizing constants and reweighing mixtures in Markov chain Monte Carlo. Technical report, Department of Statistics, University of Minnesota.

Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J. M., and Chopin, N. (2015). On particle methods for parameter estimation in general state-space models. *Statistical Science*, 30(3):328–351.

Kitagawa, G. (1998). A self-organizing state-space model. *Journal of the American Statistical Association*, 93:1203–1215.

Liu, J. S. and West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In Doucet, A., Freitas, N., and Gordon, N., editors, *Sequential Monte Carlo methods in practice*, Information Science and Statistics, pages 197–223. Springer, New York.

Maroulas, V. and Nebenfuhr, A. (2015). Tracking rapid intracellular movements: a Bayesian random set approach. *Annals of Applied Statistics*, 9(2):926–949.

Paci, L., Gelfand, A. E., and Holland, D. M. (2013). Spatio-temporal modeling for real-time ozone forecasting. *Spatial Statistics*, 4:79–93.

Ren, G., Maroulas, V., and Schizas, I. (2015). Distributed sensors-targets spatiotemporal association and tracking. *Aerospace and Electronic Systems, IEEE Transactions on*, 51(4):2570–2589.

Ren, G., Maroulas, V., and Schizas, I. (2016). Decentralized sparsity-based multi-source association and state tracking. *Signal Processing*, 120:627–643.

Robert, C. P. and Casella, G. (2010). *Introducing Monte Carlo Methods with R*. Springer, New York.

Roy, V., Evangelou, E., and Zhu, Z. (2016). Efficient estimation and prediction for the Bayesian binary spatial model with flexible link functions. *Biometrics*, 72(1):289–298.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Shaddick, G. and Zidek, J. V. (2016). *Spatio-Temporal Methods in Environmental Epidemiology*. CRC Press, Boca Raton, FL.

Storvik, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *Signal Processing, IEEE Transactions on*, 50(2):281–289.

Wikle, C. K. and Royle, J. A. (1999). Space-time dynamic design of environmental monitoring networks. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 489–507.

Sequential Empirical Bayes Method for Filtering Dynamic Spatiotemporal Processes
Web-based Supplementary Materials

Evangelos Evangelou[1]   and Vasileios Maroulas[2]

[1] Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK.
[2] Department of Mathematics, University of Tennessee, Knoxville, TN 37996, USA.

Simulation Results

The general setup of our simulations is as follows. The spatial dimension is the closed interval $[0, 1]$ and the spatial sampling locations consist of $n = 11$ equidistant points covering the spatial domain. The final sampling time is denoted by T. The latent spatiotemporal process $\mathbf{x}_t$ is simulated with constant mean $\beta = 1$, autoregressive coefficient $\alpha = 0.5$, and variance $\sigma^2 = 1$. The correlation between components of $\mathbf{x}_t$ is calculated using the exponential spatial correlation function, i.e.

$$\mathrm{Corr}(\mathbf{x}_{i,t}, \mathbf{x}_{j,t}) = \exp(-d_{ij}/\phi),$$

where d_{ij} stands for the distance between the ith and jth grid point and $\phi = 0.4$ is the range parameter. At each time t we simulate a response $\mathbf{y}_t$ conditioned on the simulated $\mathbf{x}_t$ such that $y_{i,t} \sim \mathrm{Poisson}(\tau e^{x_{i,t}})$ independently for each i, for given τ.

For inference, the priors specified in (3) were used with $a_0 = 0$, $s_0 = 0.1$, $b_0 = 0$, $q_0 = 0.01$, $c_0 = 3$, and $r_0 = 1/3$. The fine grid Φ consisted of $J = 41$ equidistant points between $\phi/2$ and 2ϕ, i.e. $\Phi = \{0.200, 0.215, 0.230, \ldots, 0.800\}$ and the coarse grid to $\Phi_K = \{0.230, 0.335, 0.440, 0.545, 0.650, 0.755\}$. The first element of Φ_K corresponds to $\tilde{\phi}$.

Algorithm 2 was run with Monte-Carlo sizes $L_k = L$ for $k = 1, \ldots, K$ and Algorithm 1 with Gibbs iterations L_g and particle size N.

1   Effect of the proposal distribution

In this section we compare the three choices of the proposal distribution discussed in the paper: (a) the Gaussian proposal; (b) the copula mean-only skewness correction; and (c) the copula mean-and-skewness correction.

The time dimension was $T = 100$. We performed 30 simulations from the model with $\tau = 1$. This model choice ensures that there is a substantial amount of skewness in the observations and will make the comparison between the three proposals more apparent. We measure the skewness of the approximation by computing the parameter $\delta^2 = a^2/(1 + a^2)$ such that values of δ^2 close to 1 give large skewness and values close to 0 give low skewness. In our simulations, the skew-normal parameter δ^2 had an average value of 0.12 with the largest value being about 0.65.

The Monte-Carlo sizes were $L = 100$, $N = 100$, and $L_g = 50$.

For each time iteration we compute the effective sample size (ESS) for each method. Ideally we want ESS to be close to N which will indicate that the proposal distribution generates good samples while a very low ESS would indicate degeneracy in the particles, which is not uncommon
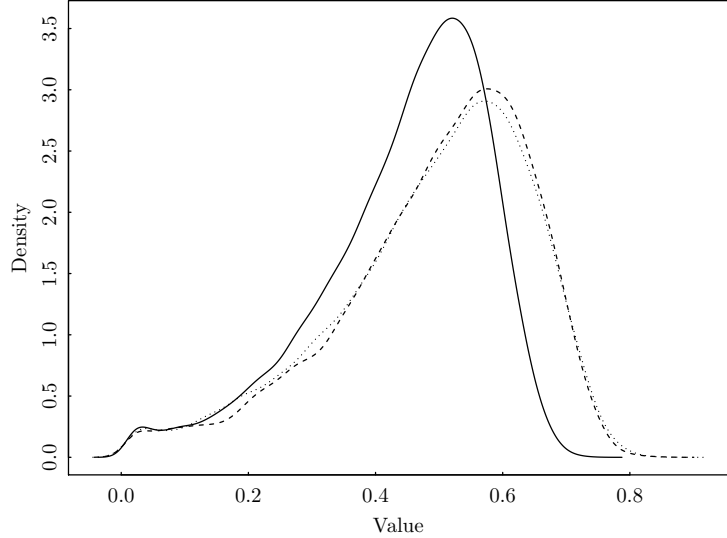
Figure 1: Density plots of the average ESS divided by the total number of generaged particles for: Gausian proposal (solid); Gaussian with mean-only correction (dashed); copula with mean-and-skewness correction (dotted).

in high dimensions. Figure 1 shows a density plot for the distribution of the average ESS over the L samples at each time iteration and for the 30 simulations (i.e, $30 \times T$ values), expressed as a proportion of the total number of samples N for each of the three proposal distributions. As shown in the figure, the uncorrected Gaussian proposal has a significantly lower ESS that the two corrected methods but the two skewness correction methods are very similar. Based on our results, and in the following, we consider the mean-only corrected proposal only.

2   Comparison with the simplified Bayes factor estimator

The simplified Bayes factor estimator is given in (10). This estimator simulates conditioned on $\phi = \tilde{\phi}$ only and uses these samples to compute the Bayes factor estimate for all $\phi \in \Phi$. In this case the reverse logistic estimates are not needed. However, as we discuss in Remark 1, this can potentially introduce bias if the true ϕ is far from $\tilde{\phi}$.

In this section we compare the bias of the simplified Bayes factor estimator with the proposed estimator (12) for the same Poisson model used in Section 1 but with increased $L = 500$ and $\tau = 10$. We consider (10) with three different values of $\tilde{\phi} = 0.395, 0.500, 0.710$, where the first value is very close to the true ϕ, the second value is at the middle of the range of Φ, and the third value is far from the true. The simplified Bayes factor estimator was tested on 30 simulated cases with burn-in $L_{\mathrm{bi}} = 50$, thinning $L_{\mathrm{th}} = 10$ and final sample size 3000. The average estimate over the 30 cases for each method was computed for each time point. This is plotted in Figure 2 for the parameters σ^2 and ϕ. The estimation for the parameters α and β did not show any obvious discrepancy.

Our results verify that the simplified Bayes factor estimator is biased and this is more apparent when $\tilde{\phi}$ is far from the true ϕ. Although the estimation for ϕ and σ^2 is biased, this does not seem to influence the estimation of α and β. This phenomenon has been observed elsewhere in the literature for the spatial-only case (see Zhang, 2002). Based on our results, the mixed Bayes factor estimator is recommended instead of the simplified one.
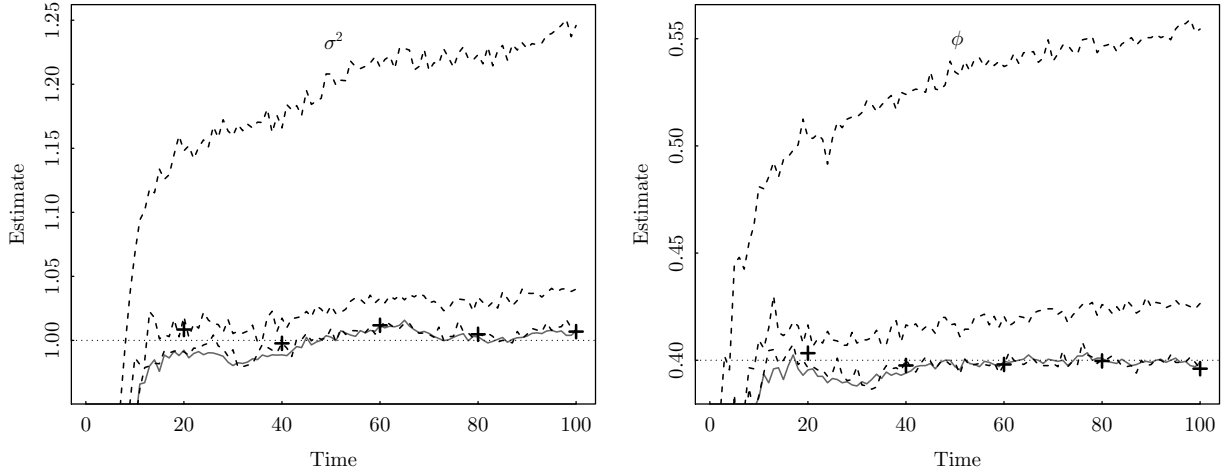
Figure 2: Average estimates for the parameters σ^2 and ϕ from the different Bayes factor estimators. The solid line is the mixed Bayes factor estimator (12). The dashed lines correspond to the simplified Bayes factor estimator (10) for $\tilde{\phi} = 0.395, 0.500, 0.710$ from bottom to top. The average offline estimates are shown by + and the true parameter value is shown by a horizontal line.

3 Estimation performance

In this section we assess the estimation performance of the proposed algorithm with the mean-only corrected proposal. We use the same setting as in Section 2 and the same 30 simulated data. We compare our estimates against an offline MCMC algorithm with the same Monte-Carlo sizes as Section 2. The offline MCMC algorithm uses Gibbs sampling for the parameters θ and a Metropolis-Hastings step to update ϕ and $x_{i,t}$, the ith component of $\mathbf{x}_t$ conditioned on everything else. The prior for ϕ was the exponential distribution with mean 0.4 and the Metropolis-Hastings step was selected for acceptance between 0.2 to 0.4. Convergence diagnostics of the MCMC output did not indicate any issues.

Because of the increasing computational time, we only ran the offline algorithm for selected time points $T_i = 20, 40, 60, 80, 100$, where at each time only data up to T_i were observed to make the results comparable with the online method.

In Figure 3 we plot the parameter estimates for each parameter in time for the online algorithm for each simulation and the distribution of the offline estimates from all simulations at the selected time points. It can be seen that the distributions from the two methods are very similar. In particular, the variability of our estimates reduces as we see more data and the bias is reduced which is a desirable property.

For each time iteration, the computing time for the sequential empirical Bayes algorithm was recorded, i.e. one iteration of Algorithm 2, and the average over the 30 simulations was taken. The average computing time is shown in Figure 4. It can be seen that the computing time does not increase in time as one would expect from an online algorithm.

Subsequently, the number of simulations was increased to 100, but in this case only the online algorithm was computed. This was to assess any potential bias in our method. The results from these simulations are shown in Figure 5 which show no apparent bias. On average, across all simulations, the four parameters are estimated accurately. Note the convergence of the estimates towards the true value and the reduction of uncertainty as more data are observed which demonstrates the suitability of our method. The sample variance of our estimates across simulations at each time
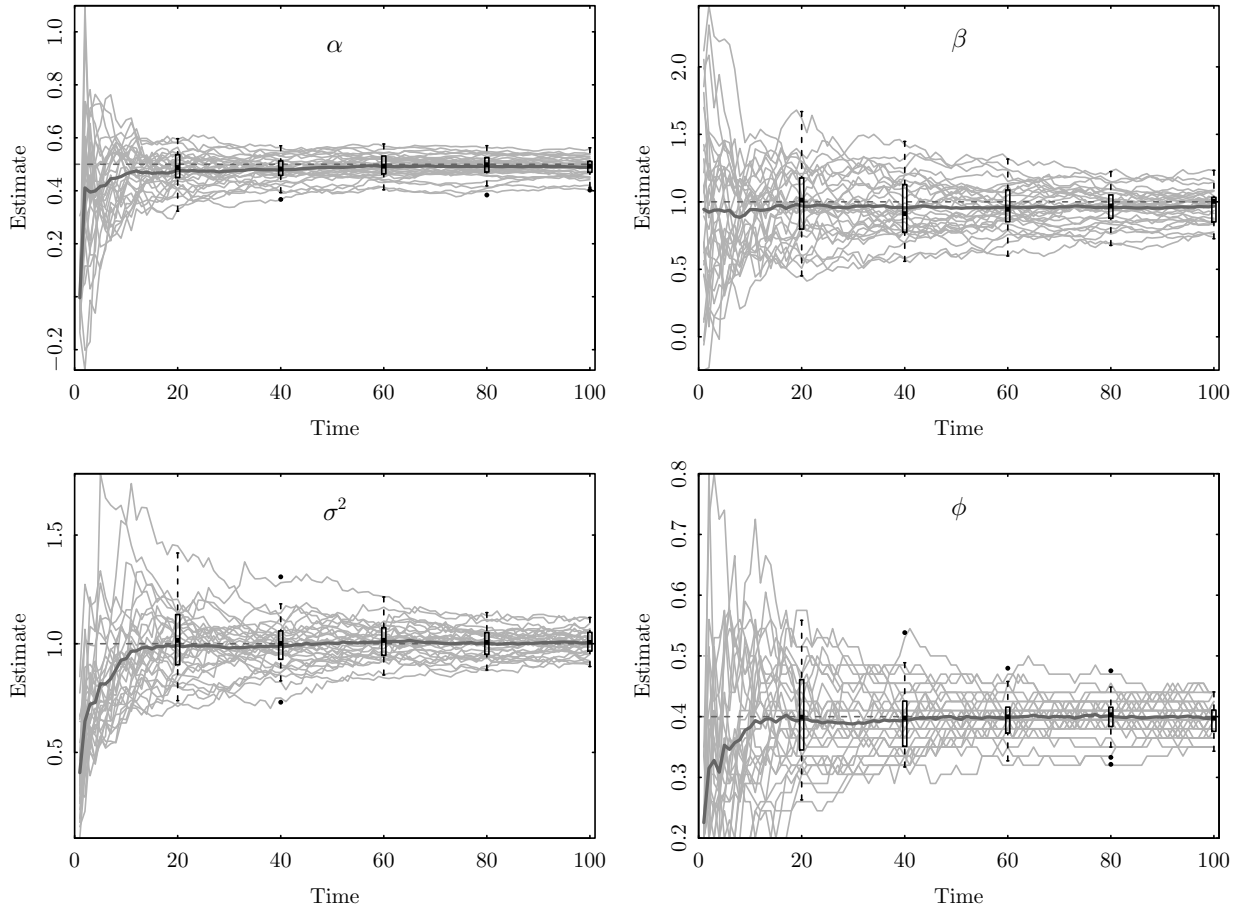
3

Figure 3: Comparison of estimates between the proposed online algorithm and the offline MCMC algorithm for the four parameters α, β, σ^2, ϕ. The light gray lines show the estimate for every simulation and the dark gray is the mean over all simulations. The boxplots show the distribution of the estimates using an offline MCMC algorithm with data available up to that time. The true parameter value is shown by a dashed line.
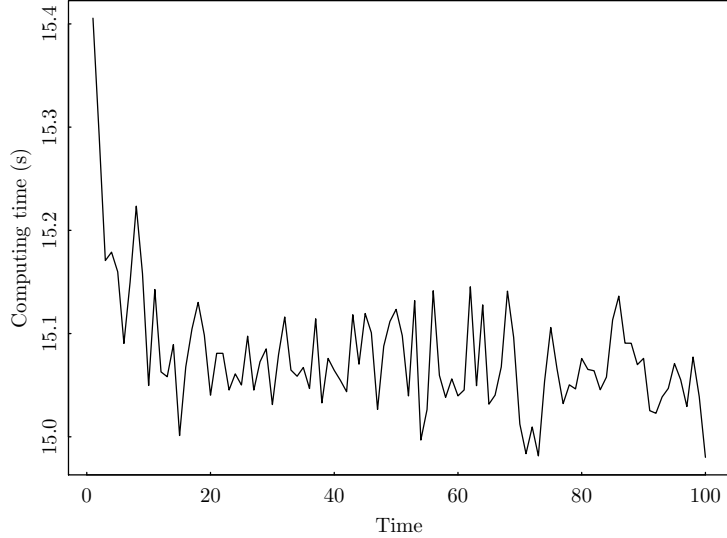
Figure 4: Average computing time per time iteration for the sequential empirical Bayes algorithm across simulations. The computations were performed on computer with Intel Core i5-2500 3.30GHz CPU and 4Gb RAM.

point was calculated and its reciprocal was plotted against time. The plots corresponding to the four parameters are shown in Figure 6. It can be seen that the variability decreases linearly which indicates a reduction in the length of the posterior credible interval in the order of $1/\sqrt{t}$ as time t increases.

4   Simulation with longer time span

In this section we use simulated data to compare the proposed algorithm against an offline MCMC algorithm. In this example the data were simulated from the model of Section 2 but with final time increased to $T = 1000$. Only one sample was generated in this example. The data were subsequently fitted using the proposed online algorithm and an offline MCMC smoothing algorithm. The priors for both methods were the same as in Section 1 and a Monte Carlo sizes were as in Section 2.

Figure 7 shows the function $\log B_{1:t}(\phi; \tilde{\phi})$ computed by the proposed online algorithm for selected values of t, along with the grids Φ_K and Φ. The maximizer of this function is the estimate for ϕ at time t. Note that, as t increases, the maximum of this function converges to the true value and the uncertainty is reduced. To derive a confidence interval we view $B_{1:t}(\phi; \tilde{\phi})$ as an unnormalized posterior pdf for ϕ and the corresponding cumulative sum is the unnormalized cumulative distribution function (cdf). We then approximate the corresponding quantiles by polynomial interpolation of ϕ against the normalized cdf.

The estimates (MC average for θ, EB estimate for ϕ) and 99% credible intervals (MC quantiles for θ, polynomial interpolation for ϕ) for each parameter using data $\mathbf{y}_{1:t}$ across t are plotted in Figure 8. As shown in the figure, since both algorithms sample from the same posterior distribution, conditioned on $\mathbf{y}_{1:t}$, the estimates and credible intervals obtained between them are very similar and capture the true parameter values even for a longer time span.
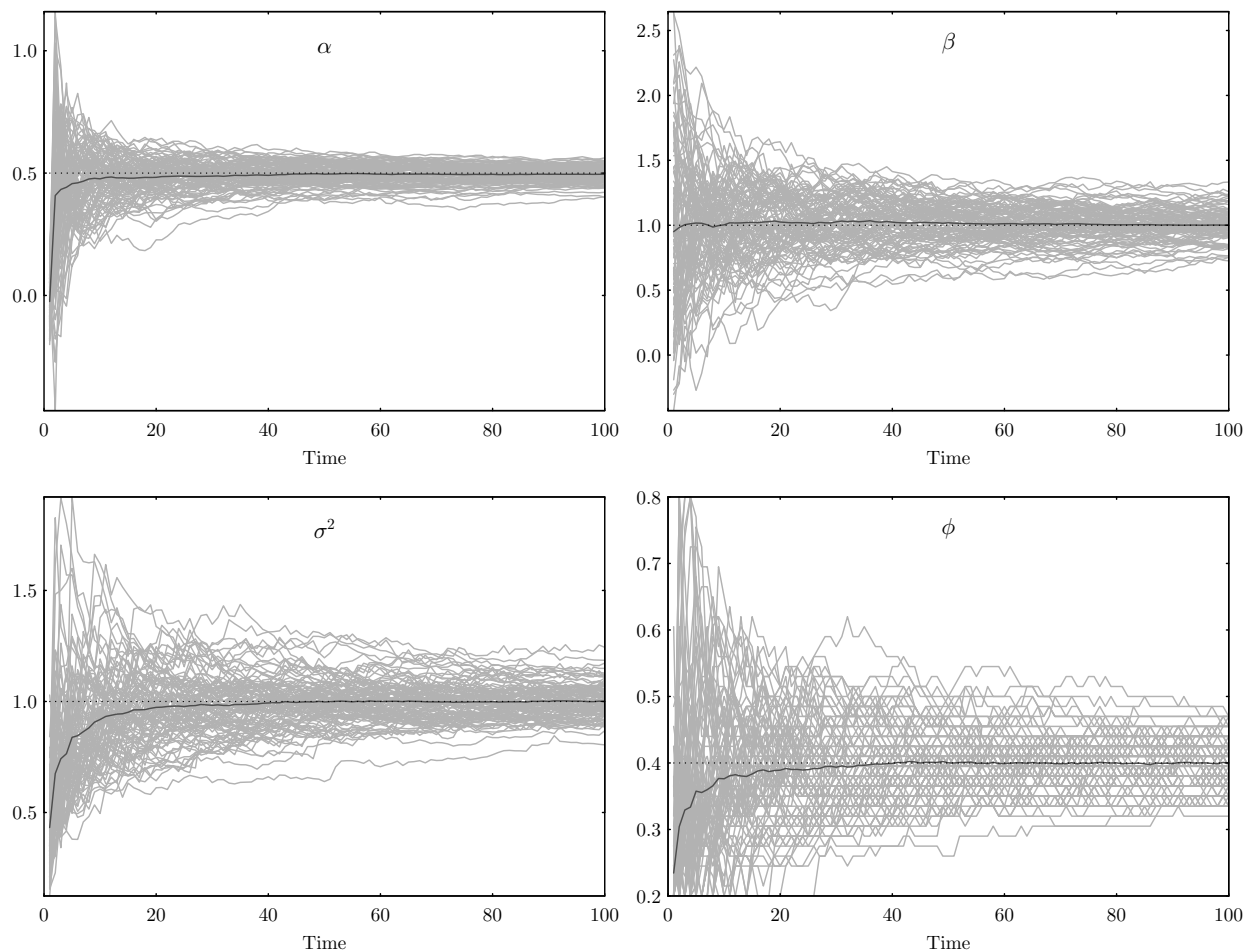
5

Figure 5: Estimates of α, β, σ^2 and ϕ in that order across time. The dotted line shows the true parameter value, and the light gray lines represent the estimates corresponding to each simulation across time. The mean across all simulations is shown by a dark gray line.

References

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics*, 58(1):129–136.
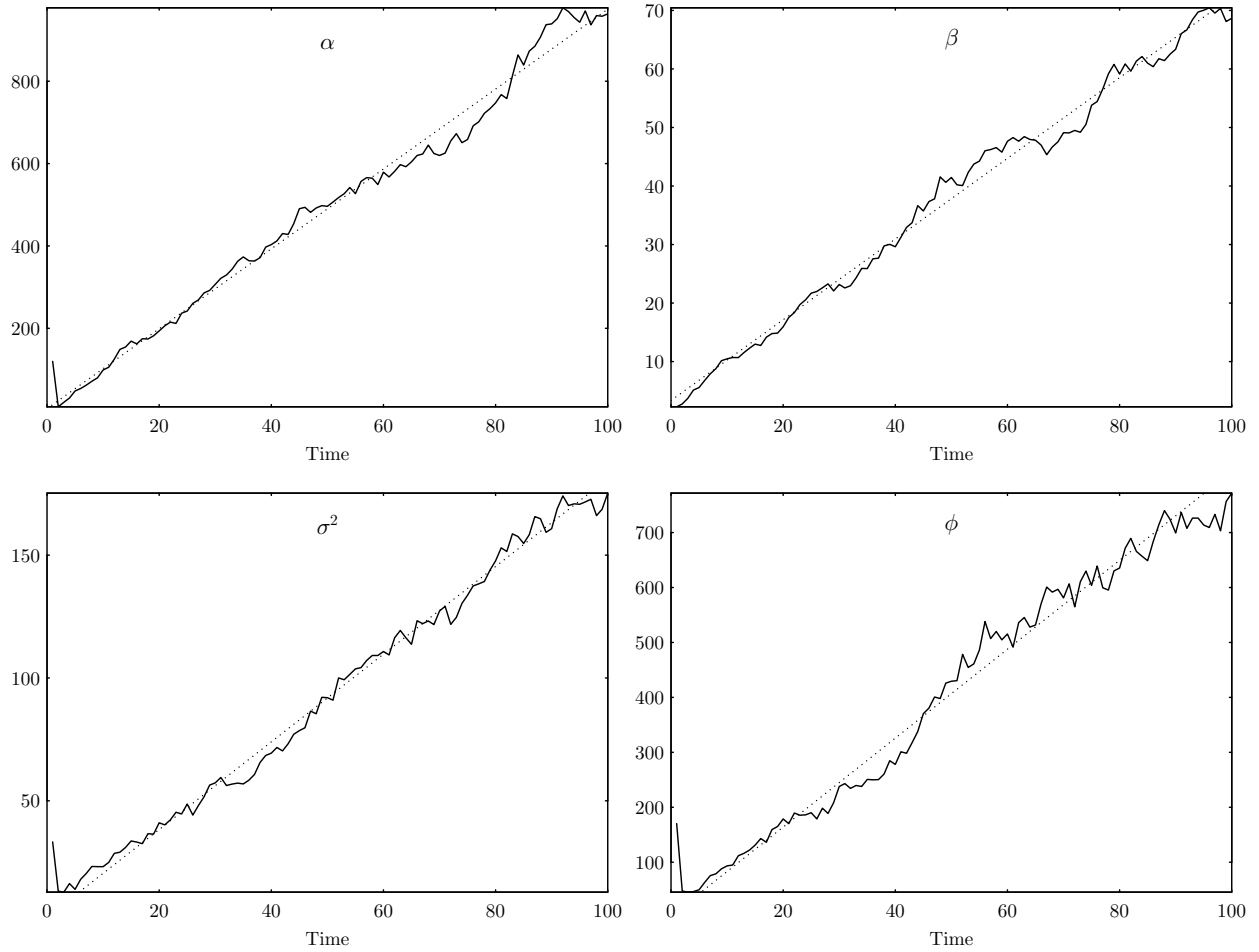
Figure 6: Reciprocal of the sample variances across simulations of the estimates for α, β, σ^2 and ϕ in that order across time. The dotted line indicates the linear rate increase in time for the reciprocal of the sample variance.

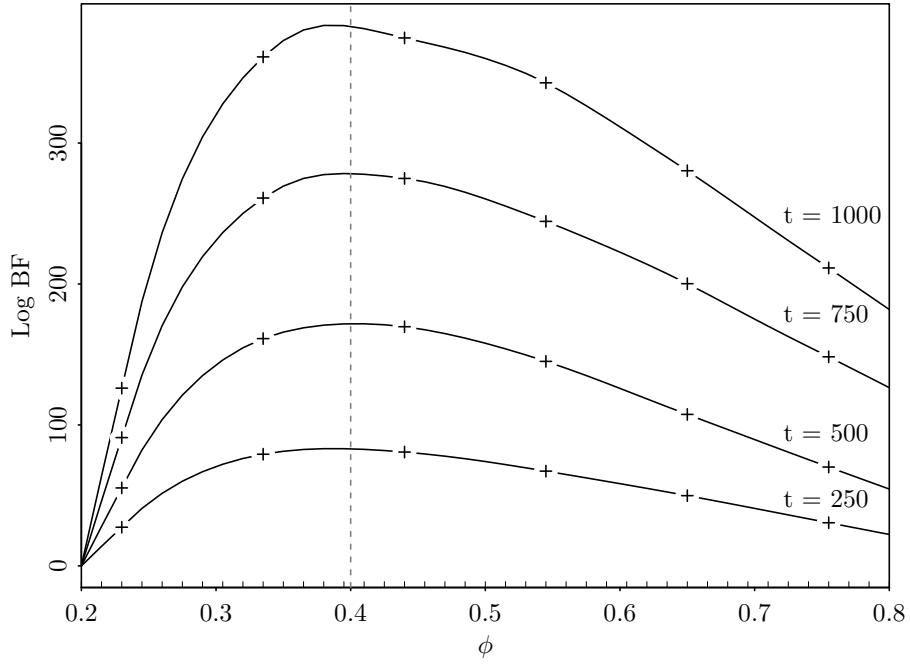Figure 7: Logarithm of the Bayes factor, $\log B_t(\phi; \tilde{\phi})$, plotted against ϕ for different times. The true ϕ is shown by a vertical line. The coarse grid Φ_K is marked by + on each line and the lines at the bottom show the fine grid Φ.
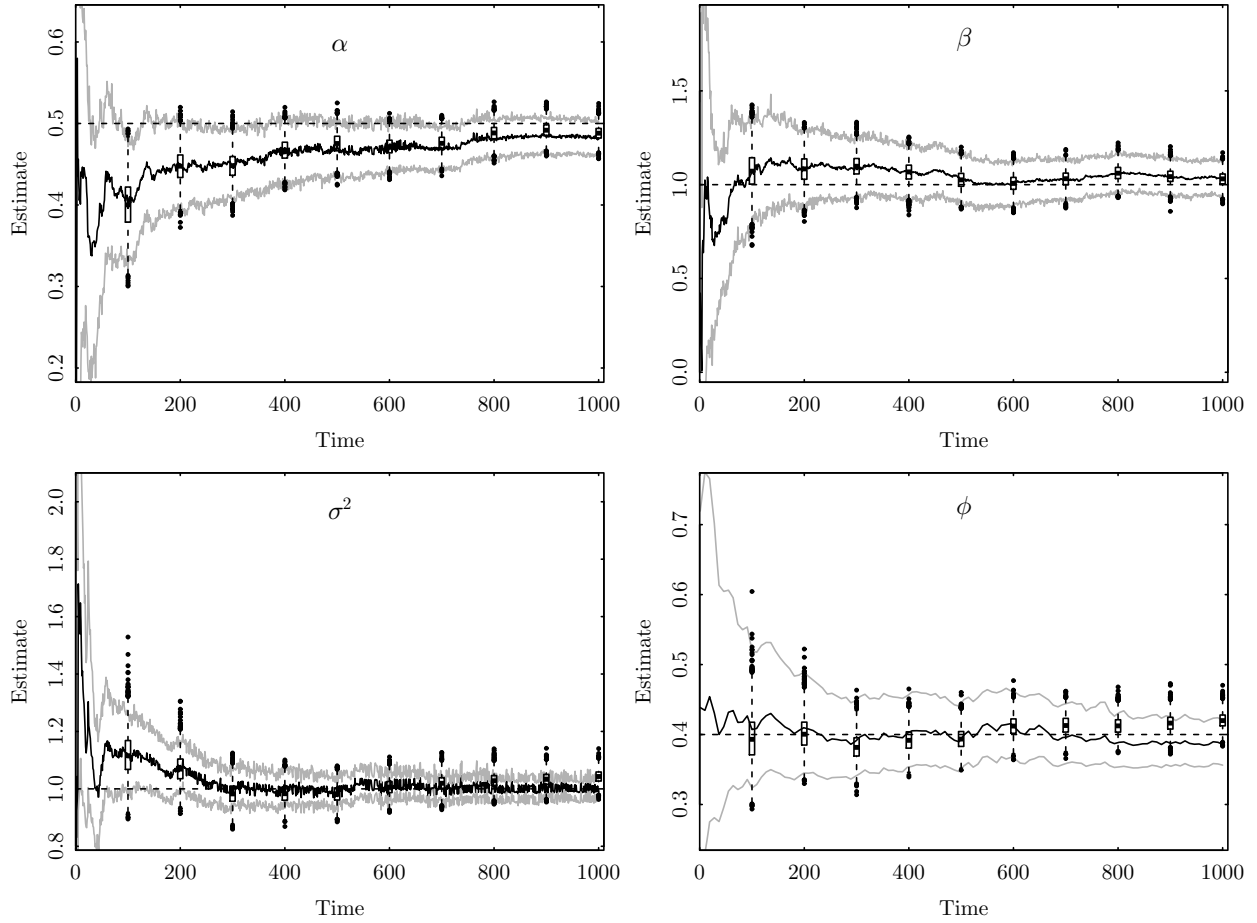
Figure 8: Parameter estimates across time using the proposed online algorithm (black lines) with 99% confidence intervals (light grey lines), and an offline algorithm (boxplots). The true parameter value is shown by a horizontal line.