Biostatistics (2013), **14**, 3, *pp.* 541–555
doi:10.1093/biostatistics/kxs052
Advance Access publication on January 4, 2013

Capturing the time-varying drivers of an epidemic using stochastic dynamical systems

JOSEPH DUREAU*, KONSTANTINOS KALOGEROPOULOS

Department of Statistics, London School of Economics and Political Science, London WC2A 2AE, UK
j.dureau@lse.ac.uk

MARC BAGUELIN

Immunisation, Hepatitis and Blood Safety Department, Health Protection Agency, Centre for the Mathematical Modeling of Infectious Diseases, London School of Hygiene and Tropical Medicine, London NW9 5EQ, UK

Summary

Epidemics are often modeled using non-linear dynamical systems observed through partial and noisy data. In this paper, we consider stochastic extensions in order to capture unknown influences (changing behaviors, public interventions, seasonal effects, etc.). These models assign diffusion processes to the time-varying parameters, and our inferential procedure is based on a suitably adjusted adaptive particle Markov chain Monte Carlo algorithm. The performance of the proposed computational methods is validated on simulated data and the adopted model is applied to the 2009 H1N1 pandemic in England. In addition to estimating the effective contact rate trajectories, the methodology is applied in real time to provide evidence in related public health decisions. Diffusion-driven susceptible exposed infected retired-type models with age structure are also introduced.

Keywords: Bayesian inference; Particle MCMC; Population epidemic model; Time-varying parameters.

1. Introduction

Epidemic models are often used to simulate disease transmission dynamics, to detect emerging outbreaks (Unkel *and others*, 2012), and to assess public health interventions (Boily *and others*, 2007). In order to capture the dynamics of epidemics, the main focus is generally made on their intrinsically dynamic elements such as the depletion of susceptibles or the population immunity evolution. Nevertheless, there are time-varying extrinsic factors that are crucial to the epidemic course. These may include social cycles (holidays), public interventions, and climatic variations. This has been illustrated for diseases such as cholera, malaria (Cazelles *and others*, 2005; Ionides *and others*, 2006), or influenza (Shaman and Kohn, 2009). These studies were conducted either by relating climatic and incidence time-series (Cazelles *and others*, 2005), which does not disentangle the effect of intrinsic and extrinsic factors, or by experimentally assessing the virus resistance in different climatic conditions (Shaman and Kohn, 2009) requiring an

*To whom correspondence should be addressed.

extrapolation to the population scale. Overall, the time-varying nature of epidemics poses a challenging statistical problem stressing the need for suitable computational tools (Ferguson, 2007).

This paper considers a flexible modeling framework that encompasses time-varying aspects of the epidemic via stochastic differential equations. We aim at providing robust inferential procedures, incorporating the uncertainty associated with key parameters and accounting for data and model limitations. In order to provide an accurate and feasible computational toolbox, we provide Markov chain Monte Carlo (MCMC) algorithms utilizing recent developments such as particle MCMC (PMCMC) algorithms (Andrieu *and others*, 2010) and adaptive techniques (Roberts and Rosenthal, 2009). Modeling aspects are presented in Section 2, while the computational framework is presented in Section 3. In Section 4, we evaluate the performance of the proposed adaptive PMCMC schemes on simulated data. In Section 5, we present various applications of the methodology to the 2009 A/H1N1 pandemic, and conclude, in Section 6, with some relevant discussion. Further simulations can be found in supplementary materials available at *Biostatistics* online.

2. MODELLING FRAMEWORK

2.1 *Epidemic models with time-varying coefficients*

We adopt a SEIR (Susceptible-Exposed-Infected-Removed) model as a guide in this paper, although the methodology can be applied to other dynamical systems. The model is set in (2.1); S accounts for susceptible, E for infected but not infective, I for infective, and R for removed individuals. New infections occur at a rate $\beta S_t(I_t/N)$, implying that the susceptible individuals make effective contacts at rate β (the effective contact rate), and only a fraction I_t/N of these contacts are made with infective individuals. The average period spent in compartments E and I is given by k^{-1} and γ^{-1}, respectively.

$$\frac{\mathrm{d}S_t}{\mathrm{d}t} = -\beta S_t \frac{I_t}{N}, \quad \frac{\mathrm{d}E_t}{\mathrm{d}t} = \beta S_t \frac{I_t}{N} - kE_t, \quad \frac{\mathrm{d}I_t}{\mathrm{d}t} = kE_t - \gamma I_t, \quad \frac{\mathrm{d}R_t}{\mathrm{d}t} = \gamma I_t. \tag{2.1}$$

The basic reproduction number, R_0, represents the number of secondary infections from a primary infected individual in a fully susceptible population. A related quantity is the effective reproduction number, R_t, refers to the number of secondary cases from an infected individual at time t. R_t is a context-dependent quantity of high interest to policy makers as it indicates the possibility for the epidemic to grow ($R_t > 1$) or to decrease ($R_t < 1$) (Anderson and May, 1992).

Epidemic models can be quite detailed (including individual characteristics, geographic information, etc.) or basic, such as the SEIR model, that geographically aggregates the cases and assumes deterministic transmission processes, occurring at a given frequency each time infected and susceptible meet. The latter are easier to estimate and interpret, but are based on strong assumptions that could lead to poor inference. In this paper, we adopt stochastic extensions of the deterministic SEIR models. The additional dynamic error is likely to contain structural mis-specifications and can subsequently be explored and potentially revised. We focus on large-scale epidemics, for which random effects in transmission processes can be considered to be well-approximated deterministically (Kurtz, 1981). We adopt the paradigm that attributes the model limitations mainly to the time-varying nature of the effective contact rate, henceforth denoted as β_t, rather than to the variability in individual characteristics or in transmission processes.

An early approach to estimate R_t can be found in Fine and Clarkson (1982). It can be implemented through discrete generation models or by reconstructing the chain of transmission (Cauchemez *and others*, 2006; Griffin *and others*, 2011). However, as R_t estimates contain both the effects of evolving transmissibility and immunity, quantitative conclusions can hardly be generalized to situations where the immunological situation is different. We therefore concentrate on estimating β_t rather than R_t. A number of approaches

use a finite dimension function space for the trajectory of β_t. Low-dimensional examples can be found in Cauchemez *and others* (2008), in which β_t is modeled as a piece-wise linear function. In some higher-complexity models, as in Cauchemez and Ferguson (2008) and Ionides *and others* (2006), β_t is estimated freely with a few-weeks resolutions. Loosely speaking, as the number of parameters for the trajectory of β_t increases, model-induced biases fade out at the expense of the variance. A compromise is required to improve robustness and is often controlled through a regularizing parameter. For example, in He *and others* (2011), β_t is estimated using cubic splines, calibrated via akaike information criterion.

2.2 *Diffusion driven epidemic models*

We consider models where diffusion processes are used for some of the coefficients in (2.1). Although alternative formulations are possible, as discussed in Section 2.1, we focus on β_t to obtain

$$\begin{cases} \dfrac{\mathrm{d}S_t}{\mathrm{d}t} = -\beta_t S_t \dfrac{I_t}{N}, \quad \dfrac{\mathrm{d}E_t}{\mathrm{d}t} = \beta_t S_t \dfrac{I_t}{N} - kE_t, \quad \dfrac{\mathrm{d}I_t}{\mathrm{d}t} = kE_t - \gamma I_t, \quad \dfrac{\mathrm{d}R_t}{\mathrm{d}t} = \gamma I_t, \\ \mathrm{d}x_t = \mu_x(x_t, \theta_x)\,\mathrm{d}t + \sigma_x(x_t, \theta_x)\,\mathrm{d}B_t, \quad x_t = h(\beta_t), \end{cases} \tag{2.2}$$

where $\mu_x(\cdot)$ denotes the drift, $\sigma_x(\cdot)$ the volatility, and $h(\cdot)$ is a positive-valued function. The assigned diffusion may capture features such as behavior changes, preventive measures, seasonal effects, holidays, etc. When prior knowledge on β_t is available, it can be reflected in $\mu_x(\cdot)$ and $\sigma_x(\cdot)$; e.g. if the contact rate is expected to converge, an Ornstein Uhlenbeck process can be chosen. Other options may include a sigmoid or a sinusoidal form; see, for example, Rasmussen *and others* (2011). In absence of prior information or when the researcher wants to impose little restrictions, a Brownian motion (BM) can be used, with $\mu_x(\cdot) \equiv 0$ and $\sigma_x(\cdot) \equiv \sigma$ (i.e. $\theta_x = \sigma$). This model, with $h(\cdot) \equiv \log(\cdot)$, is henceforth denoted as BM. The obtained output can either be reported or used as an exploratory tool to construct a more structured model; see Section 5.3 for an application. The choice of BM implies a continuous, yet non-differentiable, path satisfying the Markov property. In cases where β_t is believed to evolve as a smooth function in time, higher-order Brownian motions could be used. Loosely speaking, these may be regarded as equivalent to non-parametric approaches such as cubic splines (Wahba, 1990), with the model in (2.2) imposing a prior on β_t and σ being a regularizing factor. The rate β_t can be perceived as a product of a smooth and a rough component; the former being a population average of the intrinsic transmission procedure and latter containing extrinsic factors such as the amount of contact among individuals. It is therefore important to build a framework that contains both smooth and rough models.

The above model can be estimated with an extended Kalman filter (EKF), as in Cazelles and Chau (1997). EKF allows for fast computations, but is based on Taylor and Gaussian approximations whose error could be non-negligible; see supplementary materials (available at *Biostatistics* online) for a relevant simulation experiment. Nevertheless, the EKF can still be used as a tool to construct efficient proposal distributions for MCMC schemes. It can also be used to optimize sequential Monte Carlo (SMC) algorithms, but either at a strong computational cost (Särkkä and Sottinen, 2008) or crude time discretizations (Dukic *and others*, 2009). Next, we develop a general framework for efficient MCMC schemes that allow for good approximations.

3. Data augmentation via MCMC for diffusion driven epidemic models

This section presents a general inferential framework for diffusion-driven epidemic models. We adopt the Bayesian paradigm to incorporate parameter uncertainty and prior information in the estimates of β_t trajectories. The problem can also be cast as estimating partially observed hypoelliptic diffusions, thus presenting various difficulties (Pokern *and others*, 2009). We begin by setting the model and justifying the

need for data augmentation. Existing MCMC algorithms are considered but they can lead to extremely inefficient MCMC chains. We address the issue by taking advantage of the specific model structure to construct adaptive PMCMC schemes.

3.1 *Model and data augmentation setup*

For ease of exposition, we focus on models satisfying (2.2), but the framework covers models with different ODE systems or more time-varying coefficients, as in Section 5.3. Being in continuous time, t can take any value between t_0 and t_n. We denote the path of the ODE states vector $V_t = \{S_t, E_t, I_t, R_t\}$ between observation times t_i and t_j by $V_{i:j}$. The data, $y_{1:n} = \{y_{t_1}, \ldots, y_{t_n}\}$, usually provide information for I_t at specific times (prevalence data) or for integrals of V_t (incidence data). In either case, we assume that they are obtained with error as the collection procedure is typically associated with additional uncertainty. The noise distribution is denoted with $\mathbb{P}_y$ with density $f(y_{1:n}|V_{0:n}, \theta_y)$. Note that, in the model of (2.2), V_t can be written as a deterministic function, $g(\cdot)$, of x_t and the parameters $\theta_v = (k, \gamma, V_0)$. This function is the solution of the ODE and can be written as an intractable time integral involving x_t. Hence, the model becomes

$$\begin{cases} \mathrm{d}x_t = \mu_x(x_t, \theta_x)\,\mathrm{d}t + \sigma_x(x_t, \theta_x)\,\mathrm{d}B_t, \\ y_{1:n}|V_{0:n}, \theta_y \sim \mathbb{P}_y(y_{1:n}|V_{0:n}, \theta_y), \quad V_{0:n} = g(x_{0:n}, \theta_v). \end{cases} \tag{3.1}$$

Denote with $\mathbb{P}_x$ the distribution of the diffusion x_t defined from the stochastic differential equation above. We require the existence of a unique weak solution which translates into some mild assumptions on $\mu_x(.)$ and $\sigma_x(.)$; e.g. locally Lipschitz with a linear growth bound; see, for example, Øksendal (2003). The distribution of $\mathbb{P}_x$ may also be viewed as a prior on x_t, or else β_t. The model can now be defined from $\mathbb{P}_y$, $\mathbb{P}_x$, and the assigned priors on $\theta = \{\theta_y, \theta_v, \theta_x\}$, denoted by $\pi(\theta)$

$$\pi(x_{0:n}, \theta|y_{1:n}) \propto f(y_{1:n}|V_{0:n}, \theta_y) \times \mathrm{d}\mathbb{P}_x \times \pi(\theta). \tag{3.2}$$

Given direct observations on x_t, it would have been possible to draw approximation-free inference on $d\mathbb{P}_x$ using the approach of Beskos *and others* (2006). However, this is not possible in our case given the non-linear functionals in $g(\cdot)$ that render (3.4) intractable. We proceed by discretizing the path of x_t, and therefore of β_t and V_t. More specifically, we introduce m points between each pair of successive observation times t_i and t_{i+1} ($i = 0, 1, \ldots, n-1$). When referring to the discrete representation of a path, the superscript dis will be used; for example, for a step $\delta = 1/(m+1)$, the discrete skeleton of x_t will be denoted by $x^{\mathrm{dis}}_{0:n} = \{x_0, x_\delta, x_{2\delta}, \ldots, x_{t_n}\}$. The presence of $x^{\mathrm{dis}}_{0:n}$ allows for approximations of (3.2) through the Euler–Maruyama scheme to evaluate $d\mathbb{P}_x$

$$\begin{cases} p(x^{\mathrm{dis}}_{\delta:n}|x_0, \theta_x) = \displaystyle\prod_{i:\, t_0 < i\delta \leqslant t_n} p(x_{i\delta}|x_{(i-1)\delta}, \theta_x), \\ x_{i\delta}|x_{(i-1)\delta} \sim \mathcal{N}\{x_{(i-1)\delta} + \delta\mu_x(x_{(i-1)\delta}, \theta_x),\ \delta\sigma_x(x_{(i-1)\delta}, \theta_x)^2\}. \end{cases} \tag{3.3}$$

Moreover, given $x^{\mathrm{dis}}_{0:n}$, the ODE can be solved numerically to obtain $V^{\mathrm{dis}}_{0:n}$ and evaluate $f(\cdot)$. The approximation error can be made arbitrarily small by increasing the user-specified parameter m.

3.2 *Data augmentation via Gibbs schemes*

Model (3.1) can be put in the context of Chib *and others* (2006), Golightly and Wilkinson (2008), or Kalogeropoulos (2007). In these approaches, a Gibbs scheme can be used to sample from the joint posterior in (3.2) of $x^{\mathrm{dis}}_{0:n}$ and θ. The data augmentation algorithm alternates between drawing $x^{\mathrm{dis}}_{0:n}$ given θ,

and updating θ conditional on the augmented path $x_{0:n}^{\text{dis}}$. The MCMC protocol ensures that the chain provides samples from the marginal posteriors of $x_{0:n}^{\text{dis}}$ and θ. Nevertheless, the properties of the algorithm may become unacceptably poor. There are two essential issues associated with such schemes. The first concerns the non-trivial step of sampling on the diffusion path space of x_t. The second problem is caused by the high posterior correlations between $x_{0:n}^{\text{dis}}$ and θ, leading to reducible chains as m increases (Roberts and Stramer, 2001).

The majority of the literature on data augmentation schemes for diffusions handles the conditional updates of $x_{0:n}^{\text{dis}}$ with an independence sampler. As it is difficult to find good proposal distributions for the entire $x_{0:n}^{\text{dis}}$, the path is usually split into blocks. Overlapping blocking strategies are essential to ensure that all points are updated and continuity of the path is retained. An alternative way of updating $x_{0:n}^{\text{dis}}$ is to use the particle filter via the particle Gibbs algorithm of Andrieu *and others* (2010). But unless the issue of high posterior correlation between $x_{0:n}^{\text{dis}}$ and θ is resolved, none of these schemes will improve the overall MCMC performance. The problem is caused by the quadratic variation process of x_t that identifies θ_x. For $\sigma_x(x_s, \theta_x) \equiv \sigma$, we obtain

$$\lim_{\delta \to 0} \sum_{i:\ t_0 < i\delta \leqslant t_n} (x_{i\delta} - x_{(i-1)\delta})^2 = \int_{t_0}^{t_n} \sigma^2 \, \mathrm{d}s = \sigma^2 (t_n - t_0). \tag{3.4}$$

Thus, the conditional posterior of σ converges to a point mass as δ tends to 0. In practice, this translates into an increasingly slow MCMC algorithm with a convergence rate of $O(m)$ (Roberts and Stramer, 2001). Schemes with a fixed m (Cori *and others*, 2009) could work in some occasions but the approximation error could be substantial. In some cases, the problem can be tackled with suitable reparameterization. The approach of Roberts and Stramer (2001) involves transforming to a diffusion $\dot{x}_t$ with unit volatility. An alternative scheme is offered by Chib *and others* (2006) where the driving BM of x_t is being used. In these algorithms the ODE states vector $V_{0:n}^{\text{dis}}$ becomes a function of σ, $\dot{x}_{0:n}$ and θ_v. Hence, in a Metropolis step, every proposed value of σ^* is associated with the corresponding values of $V_{0:n}^{\text{dis}*}$. This succeeds into breaking the perfect dependence between $V_{0:n}^{\text{dis}}$ and σ, even for $m \to \infty$. But since components of $V_{0:n}^{\text{dis}}$ (or functionals thereof) are observed with error, the associated proposed values $V_{0:n}^{\text{dis}*}$ should be close to the data for the move to be accepted. As the observation error becomes small and the data increase, this becomes increasingly difficult and leads to very small moves for σ and poor MCMC mixing. More details and simulations supporting this argument are provided in the supplementary materials (Appendix E available at *Biostatistics* online). Consequently, we overcome this issue by updating $x_{0:n}^{\text{dis}}$ and θ jointly via the PMCMC algorithm, which is essential as it is not straightforward to implement joint updates with the other approaches mentioned in this section.

3.3 *Adaptive particle MCMC algorithms*

Particle filters are SMC algorithms used to recursively explore conditional densities in state–space models (Doucet and Johansen, 2009). For given values of θ, N particles $(\tilde{x}_i^j)$ are sequentially propagated from t_0 to t_n. In various time steps t_i, the trajectories that best fit the data $y_{1:i}$ are given more weight through resampling. Algorithm 1 shows how they can be applied in our context. The quantity $L^{i+1}(\theta)$ provides unbiased estimates of $p(y_{1:i}|\theta)$ and the resampling step is essential to control the variance of that estimate over time. Algorithm 1 also provides a random sample from $p(x_{1:i}|y_{1:n}, \theta)$. In order to sample from $\pi(x_{1:n}, \theta|y_{1:n})$, the PMCMC algorithm can be used. PMCMC was introduced in Andrieu *and others* (2010) and successfully integrates particle filters in MCMC algorithms. Its implementation is presented in Algorithm 2. The issues of Section 3.2 are now addressed as $x_{0:n}^{\text{dis}}$ and θ are sampled jointly. In other words, $x_{0:n}^{\text{dis}}$ is being numerically integrated out, while a sample from its posterior is obtained at each MCMC iteration.

Algorithm 1 Particle Filter algorithm

Initialise: Set $L^0(\theta) = 1$, $W_0^j = \frac{1}{N}$, sample $(\tilde{x}_0^j)_{j=1,\ldots,N}$ from $p(x_0|\theta)$ and calculate $(\tilde{V}_0^j)_{j=1,\ldots,N}$ by solving the ODE (for example with the Euler scheme)
for $i = 0$ to $n - 1$ **do**
  for $j = 1$ to N **do**
    Sample $(\tilde{x}_{i:i+1}^j)$ from (3.3) and calculate $(\tilde{V}_{i:i+1}^j)$ by solving the ODE
    Set $\alpha^j = f(y_{i+1}|\tilde{V}_{0:i+1}^j)$
  end for
  Set $W_{i+1}^j = \frac{\alpha^j}{\sum_{k=1}^N \alpha^k}$, and $L^{i+1}(\theta) = L^i(\theta) \times \frac{1}{N}\sum \alpha^j$
  Resample $(\tilde{V}_{0:i+1}^j, \tilde{x}_{0:i+1}^j)_{j=1,\ldots,N}$ according to (W_{i+1}^j),
end for

Algorithm 2 Particle MCMC algorithm (particle Marginal Metropolis Hastings version)

Initialise: Set current θ value, $\tilde{\theta}$, to an initial value. Use Particle Smoother (PS) according to Algorithm 1 to compute $\hat{p}(y_{1:n}|\tilde{\theta}) = L(\tilde{\theta})$ and sample $\tilde{x}_{1:n}^{\tilde{\theta}}$ from $p(x_{1:n}|y_{1:n}, \tilde{\theta})$
for $It = 1$ to $NIterations$ **do**
  Sample $\tilde{\theta}^*$ from $Q(\tilde{\theta}, .)$
  Use PS to compute $L(\tilde{\theta}^*)$ and sample $\tilde{x}_{1:n}^{\tilde{\theta}^*}$ from $\hat{p}(x_{1:n}|y_{1:n}, \tilde{\theta}^*)$
  Do $\tilde{\theta} = \tilde{\theta}^*$ (and $\tilde{x}_{1:n}^{\tilde{\theta}} = \tilde{x}_{1:n}^{\tilde{\theta}^*}$) with probability $1 \wedge \frac{L(\tilde{\theta}^*)Q(\tilde{\theta}^*,\tilde{\theta})}{L(\tilde{\theta})Q(\tilde{\theta},\tilde{\theta}^*)}$
  Record $\tilde{\theta}$ and $\tilde{x}_{1:n}^{\theta}$
end for

While the PMCMC algorithm is theoretically valid even for a single particle, large values of N are usually required for reasonably stable acceptance rates and large moves in the θ space; see the supplementary materials (available at *Biostatistics* online) for a relevant simulation exercise. It is therefore essential to update the d-dimensional θ at once, marking the proposal $Q(\theta, .)$ crucial to the overall MCMC performance. In this paper, we propose to use the adaptive Metropolis algorithm of Roberts and Rosenthal (2009). After transforming the parameters to take values in the real line we use a Normal distribution centered at the current value of θ and with covariance given by $\epsilon\Sigma$. Static random walk metropolis proposals set $\Sigma = I_d$ or $\Sigma = \hat{\Sigma}$ and tune ϵ to obtain acceptance rate of 0.234. Adaptive schemes change the value ϵ for each iteration i through diminishing adaptation; e.g. by $\epsilon_{i+1} = \exp\{\log(\epsilon_i) + \alpha_1^n(\text{AccRate} - 0.234)\}$, where $\alpha_1 = 0.999$ and 'AccRate' denotes the acceptance rate up to iteration i. The covariance matrix Σ_{i+1} can also be updated as

$$\alpha_2 \mathcal{N}\left(\theta, \epsilon\frac{2.38^2}{d}\Sigma_0\right) + (1-\alpha_2)\mathcal{N}\left(\theta, \epsilon\frac{2.38^2}{d}\Sigma_i\right),$$

where α_2 is usually set to 0.05, Σ_i is the posterior covariance matrix estimated by the draws up to i and Σ_0 should be specified in advance. In this paper, we enhance the above adaptive algorithms utilizing information from the EKF to estimate the covariance $\hat{\Sigma}$ or Σ_0. One choice, EK-Mode, is the observed information matrix at the mode identified by EKF, evaluated through numerical differentiation. Another choice, EK-MCMC, is to run an approximate MCMC scheme based on the EKF approximation of the likelihood and compute the posterior covariance from the draws. Note that the computational burden of these methods is marginal with regard to the PMCMC. As demonstrated in Section 4, the use of EKF can result in substantial improvement.

4. Simulation experiments

The proposed algorithms are illustrated and tested on simulated data in this section. We focus on the BM model, where $\log(\beta_t)$ follows a Brownian motion with volatility σ, corresponding to the case of little information on the shape of β_t. The trajectories of β_t were drawn either from the BM model itself (experiment 1) or from a deterministic sigmoid curve (experiment 2). The data y_i, where $i = 1, \ldots, 50$ represent noisy observations of weekly new cases of the epidemic $\int_{\text{week } i} kE_t \, \mathrm{d}t$. We complete the model by assigning a Normal distribution to each $\log(y_i)$ with mean $\log\left(\int_{\text{week } i} kE_t \, \mathrm{d}t\right)$ and variance τ^2. The parameters were tuned to obtain realistic epidemic incidence curves, and observations were generated setting $\tau = 0.1$. The assigned priors were informative for k, γ, and $R(t_0)$ and vague for $E(t_0)$, $I(t_0)$, σ, and τ, as in Section 5.1. We used 3000 particles and 100 000 MCMC iterations after a long burn-in period. Figure 1 shows estimates and 95% point-wise credible intervals of the path, provided by the adaptive PMCMC initialized with EK-MCMC. The posterior output is in good agreement with the simulation trajectories suggesting that the underlying trajectory of β_t can be estimated reasonably well from the partial and noisy observations considered. More can be found in the supplementary materials (Appendix C available at *Biostatistics* online), where we also considered a value of $\tau = 0.05$ and obtained similar results.

Next, we use the data of experiment 1 to compare the proposed adaptive PMCMC schemes. Comparison is made in terms of the effective sample size $\text{ESS} = \left(1 + 2\sum_{i \geqslant 1} \eta(i)\right)^{-1}$, with $\sum_i \eta(i)$ being the sum of the lagged sample auto-correlations, as in Geyer (1992). We record the minimum ESS among the MCMC components and multiply by 100 to monitor the percentage of the total iterations that can be considered as independent. We consider three covariance matrices for each of the two adaptive algorithms defined in Section 3.3: I_d and the ones from EK-Mode and EK-MCMC. For the schemes that adapt ϵ the minimum ESS was 0.008% (I_d), 0.19% (EK-Mode), and 0.54%(EK-MCMC), whereas for the schemes that adapt Σ we got 0.57%, 1.24%, and 1.38%, respectively. Clearly, adapting Σ is crucial to obtain a reasonable performance, unless the matrices obtained from EK-Mode or EK-MCMC are used. The proposed adaptive algorithms induce substantial improvement that is expected to intensify as the dimension of θ increases.

5. The 2009 A/H1N1 pandemic

5.1 *Data, model, and estimates*

The proposed methodology is illustrated on data from the A/H1N1(2009) pandemic in England between June and December 2009. The data consist of estimates of weekly ILI cases $y_{1:n}$ given by the Health Protection Agency (Baguelin *and others*, 2010). The estimates were obtained from the recorded ILI cases among a selected sample of GPs. They accounted for over-reporting due to similarities in symptoms with other respiratory diseases, based on subsequent virological positivity tests. Corrections for asymptomatic infections and the patients' propensity to consult were also made. Overall the two datasets are different by a multiplicative coefficient $c = 10$, whose value is also supported by a further serological survey (Miller *and others*, 2010). In our analysis c is initially held fixed to 10, but this choice is explored further in Section 5.2. We adopt a model that admits noisy data to reflect the associated uncertainty. The noise model of Section 4 was used, combined with a BM formulation of $\mathbb{P}_x$. Vague priors, $N_{>0}(0, 10^6)$, were put on τ, σ and β_0. The priors for k and γ were obtained from additional data sources (Baguelin *and others*, 2010), the results of which are summarized through Normal distributions that place 95% probability in a symmetric manner between 1.55 and 1.63 days for the latent period k^{-1}, and between 0.93 and 1.23 days for the infectious period γ^{-1}. A Dirichlet distribution was used for the initial proportions in compartments S, E, I, R, constraining the mean of the one in R to be 0.15, its variance 0.15^2, and the means of the other initial proportions to be equal.

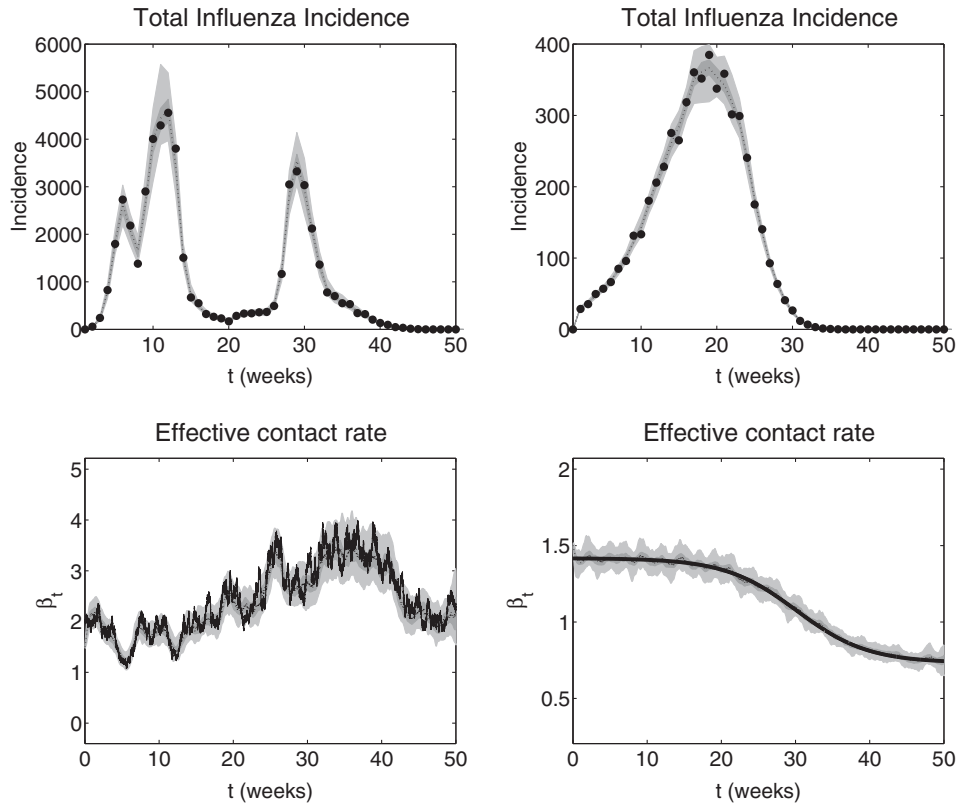

Fig. 1. Illustration of how the underlying dynamic of the effective contact rate can be estimated from weekly recorded cases. Black dots indicate simulated observed incidence (top panels). Black lines indicate simulated effective contact rate trajectories (bottom panels). Black dotted lines indicate the mean of the point-wise posterior density. Dark and light grey areas show credible intervals, respectively, at 50% and 95% levels. Top panels: simulated weekly numbers of cases observed with noise, and corresponding model-based offline reconstructions (left: experiment 1, right: experiment 2). Bottom panels: simulated and estimated trajectory of the effective contact rate (left: experiment 1, right: experiment 2).

The adaptive EK-MCMC algorithm was applied to the data and Figure 2 depicts the incidence curve together with the posterior mean and point-wise 95% credible intervals. Estimates of β_t are also displayed indicating various changes over time. The changes in β_t are consistent with the argument that schools closure for holidays have been driving the epidemic: different values are observed during school and holidays periods, appearing to be synchronized with schools opening and closing. Posterior summaries for the static parameters, as well as a sensitivity analysis on the priors can be found in the supplementary materials available at *Biostatistics* online. These suggest that inference is quite sensitive to the choice of prior for k and γ, but not for the remaining parameters. It would be interesting to repeat the procedure under an evidence synthesis framework and vague priors.

5.2 *Application in real time. Was the first wave waning due to depletion of susceptibles?*

In this section, the methodology of the paper is applied in real time, i.e. considering partial datasets from June 2009 up to the 20th of July, the 7th of September and the 26th of October. Each time the algorithm

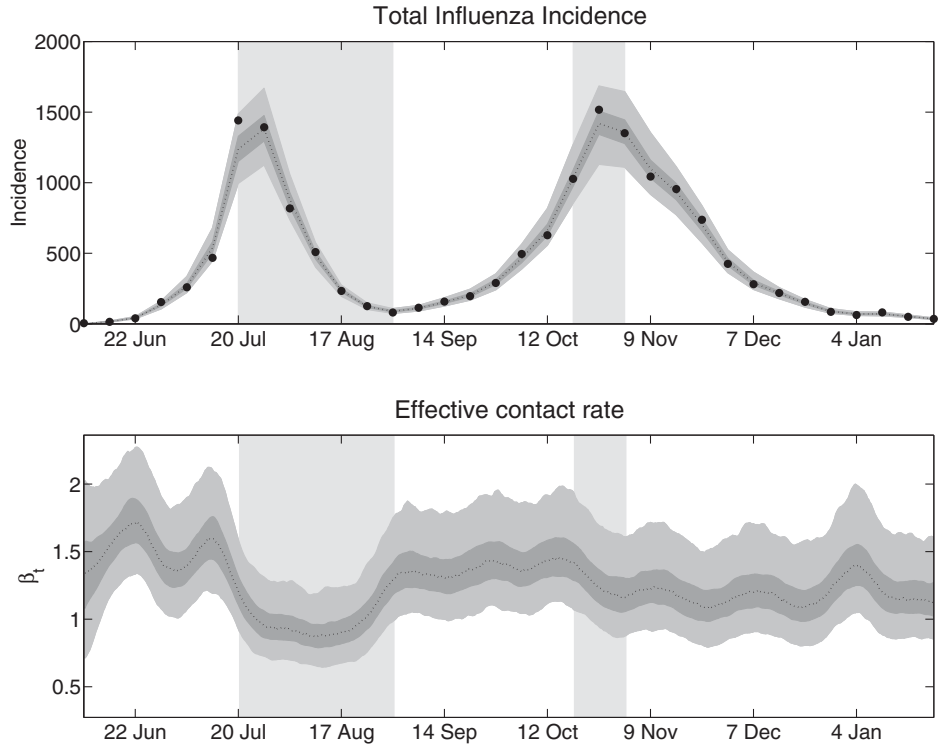

Fig. 2. Weekly incidence data from the A/H1N1 2009 influenza pandemic and corresponding offline estimates of the effective contact rate. Black dots indicate incidence estimates provided by the Health Protection Agency. Black dotted lines indicate the mean of the point-wise posterior density. Dark and light grey areas show credible intervals, respectively, at 50% and 95% levels. Holidays are indicated by a very light grey area. Top: observations of the weekly total number of A/H1N1 influenza cases in London (per 100 000 inhabs.) and model-based offline reconstruction. Bottom: offline estimates of the effective contact rate.

is run from scratch to provide samples from the joint posterior $\pi(x_{1:i}, \theta | y_{1:i})$. From a computational cost point of view this procedure can be improved further by utilizing previous MCMC runs, for example, under the SMC^2 framework (Chopin *and others*, 2011). We did not pursue this direction further, as the PMCMC algorithm runs quite fast (<2 h on a standard PC). In order to reduce uncertainty, especially at early stages, the value of τ was set to 0.1 rather than being estimated as in Section 5.1. We otherwise use the same model as before. A model with integrated BM was also fit but BM was chosen in terms of DIC; see supplementary materials (Appendix C available at *Biostatistics* online). The main results are shown in Figure 3.

On August 1st, the first wave of the epidemic had waned, incidence rates were decreasing and schools had closed. There were two competing scenarios to explain the epidemic decline: (i) holidays had caused the waning of the epidemic by lowering the effective contact rate. Hence, a similar or stronger wave could occur when schools would reopen in September in colder climatic conditions. (ii) The epidemic had stopped independently of holidays because a critical proportion of the population had been infected, conferring a sufficient level of herd immunity to stop the epidemic. In this case, no second wave was to be expected in September. On August 1st there was great uncertainty around the value of c (Baguelin *and others*, 2010), which is crucial in distinguishing between the two scenarios. We therefore conducted the following exercise.

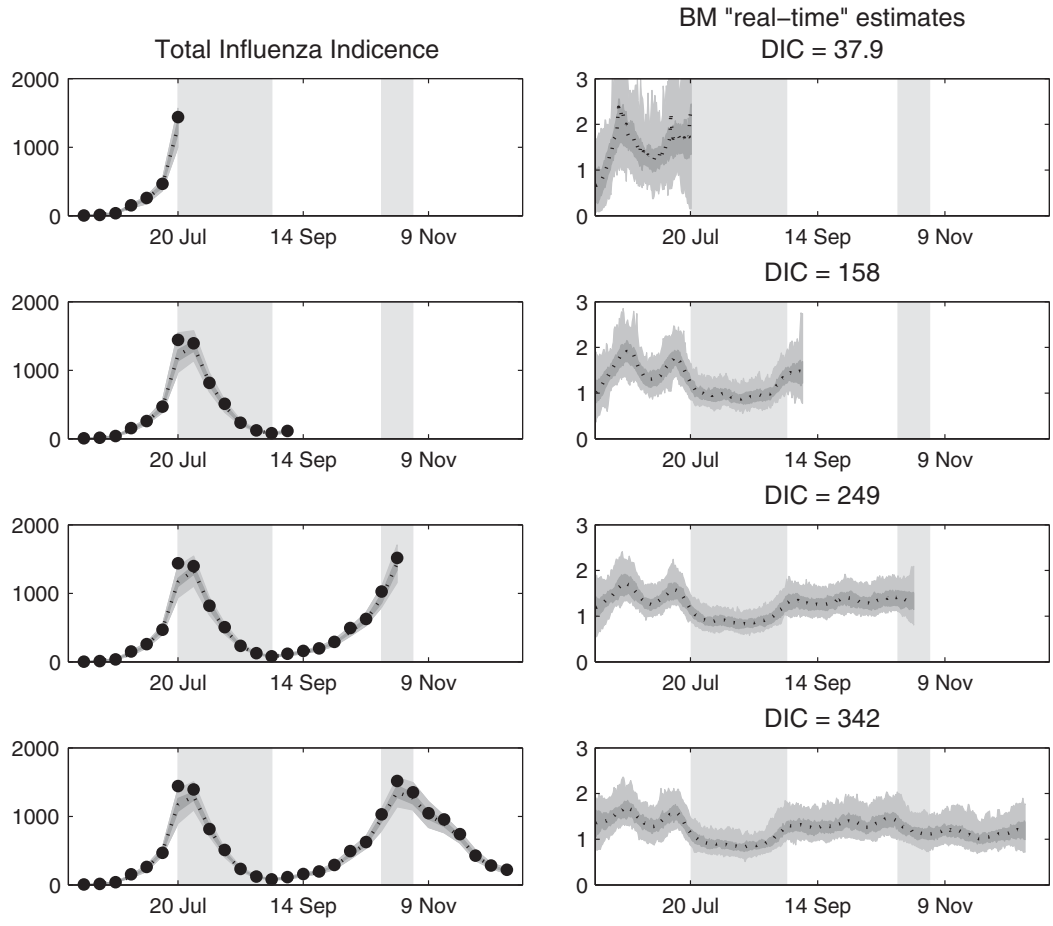

Fig. 3. What could have been inferred by carefully following the epidemic in real time? Black dots indicate observed incidence estimates provided by the Health Protection Agency (left panels). Black dotted lines indicate the mean of the point-wise posterior density. Dark and light grey areas, respectively, indicate 50% and 95% credible intervals of the posterior density. Holidays are indicated by a very light grey area. Left panels: HPA estimates of the weekly total number of A/H1N1 influenza cases in London (per 100 000 inhabs.). Right panels: "real-time" estimates of the effective contact rate.

The PMCMC algorithm, run up to August 1st, provides samples from the posterior of the difference in β_t between July 13th (before the decrease in incidence) and August 1st. For $c = 10$, the 97.5% point of this posterior is -0.32, indicating a decrease in β_t. The latter supports scenario (i), as the competing scenario is associated with a zero-decrease in β_t. Nevertheless, as this value depends on c, the algorithm was run for different values of it ranging from 20 to 150. The results appear on Figure 4. Note that the 97.5% point of interest increases as a function of c and reaches 0 for a correction factor close to 70. As this level seemed unrealistic (Baguelin *and others*, 2010), the experiment provides evidence in favor of scenario (i) highlighting the danger of a second wave in September, that actually occurred. Such evidence can be important for decision-makers, especially when considering implementations of preventive measures as vaccines.

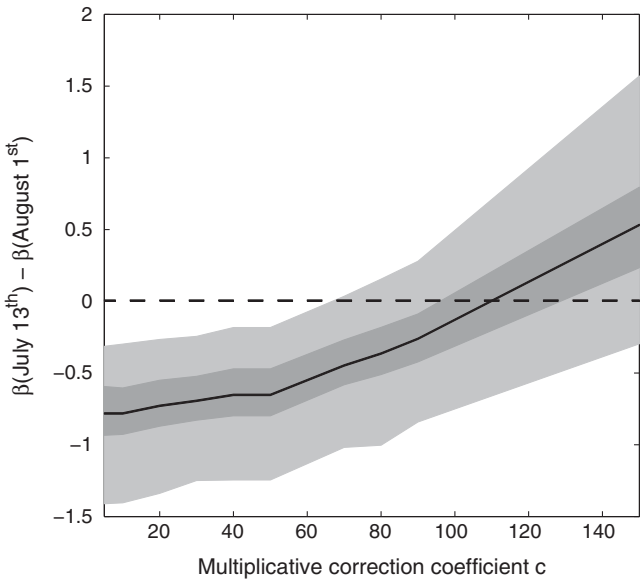

Fig. 4. The implication of different scenarios for the real value of underreporting on the decrease of the effective contact rate between July 13th and August 1st. For each value of c, the mean of the posterior density for β(August 1st)–β(July 13th) is plotted in black. Dark and light grey areas, respectively, indicate 50% and 95% credible intervals of the posterior density. The dotted line locates the scenario with no change in the effective contact rate.

5.3 *A multiple age group diffusion-driven SEIR model*

The analysis of Section 5.1 can be used to construct more structured models. For example, the effect of holidays is evident and may differ from children to adults, thus casting doubts on the assumption of a homogeneous population. It seems more natural to consider a model with two age groups (c:children and a:adults) and target all possible effective contact rates among them. In our notation, β^{ca} refers to the effective contact rate from children to adults and S^{c} denotes the number of susceptible children. For reasons of parsimony, we assign Brownian motions to $\log(\beta_t^{\mathrm{cc}})$, $\log(\beta_t^{\mathrm{aa}})$ and treat β^{ca} and β^{ac} as constants. We also set $\beta^{\mathrm{ca}} = \beta^{\mathrm{ac}} = b$ in line with various multiple age groups epidemic models in different settings (e.g. Whitaker and Farrington, 2004). The dynamic part of the model is now given by

$$\begin{cases} \dfrac{\mathrm{d}S_t^{\mathrm{c}}}{\mathrm{d}t} = -S_t^{\mathrm{c}}\left(\beta_t^{\mathrm{cc}}\dfrac{I_t^{\mathrm{c}}}{N^{\mathrm{c}}} + b\dfrac{I_t^{\mathrm{a}}}{N^{\mathrm{a}}}\right), & \dfrac{\mathrm{d}E_t^{\mathrm{c}}}{\mathrm{d}t} = S_t^{\mathrm{c}}\left(\beta_t^{\mathrm{cc}}\dfrac{I_t^{\mathrm{c}}}{N^{\mathrm{c}}} + b\dfrac{I_t^{\mathrm{a}}}{N^{\mathrm{a}}}\right) - kE_t^{\mathrm{c}}, \\ \dfrac{\mathrm{d}S_t^{a}}{\mathrm{d}t} = -S_t^{a}\left(\beta_t^{\mathrm{aa}}\dfrac{I_t^{\mathrm{a}}}{N^{\mathrm{a}}} + b\dfrac{I_t^{\mathrm{c}}}{N^{\mathrm{c}}}\right), & \dfrac{\mathrm{d}E_t^{a}}{\mathrm{d}t} = S_t^{a}\left(\beta_t^{\mathrm{aa}}\dfrac{I_t^{\mathrm{a}}}{N^{\mathrm{a}}} + b\dfrac{I_t^{\mathrm{c}}}{N^{\mathrm{c}}}\right) - kE_t^{\mathrm{a}}, \\ \dfrac{\mathrm{d}I_t^{\mathrm{c}}}{\mathrm{d}t} = kE_t^{\mathrm{c}} - \gamma I_t^{\mathrm{c}}, & \dfrac{\mathrm{d}R_t^{\mathrm{c}}}{\mathrm{d}t} = \gamma I_t^{\mathrm{c}}, \dfrac{\mathrm{d}I_t^{\mathrm{a}}}{\mathrm{d}t} = kE_t^{\mathrm{a}} - \gamma I_t^{\mathrm{a}}, \dfrac{\mathrm{d}R_t^{\mathrm{a}}}{\mathrm{d}t} = \gamma I_t^{\mathrm{a}}. \end{cases} \tag{5.1}$$

The data from the A/H1N1(2009) pandemic provide incidence estimates for children and adults separately so that they can be used to estimate the model of (5.1). If only final outcome data were available, not all effective contact rate parameters would be estimable. However, the temporal dataset provides extra information by the relative variation of susceptible and infective population in adults versus children. We applied the EK-MCMC scheme, which was essential in order to obtain reasonable MCMC performance. Figure 5 depicts the results. Unlike earlier attempts with versions of a multi-group model with a single diffusion driving all contact rates, the fit appears to be good. The trajectory of children seems to

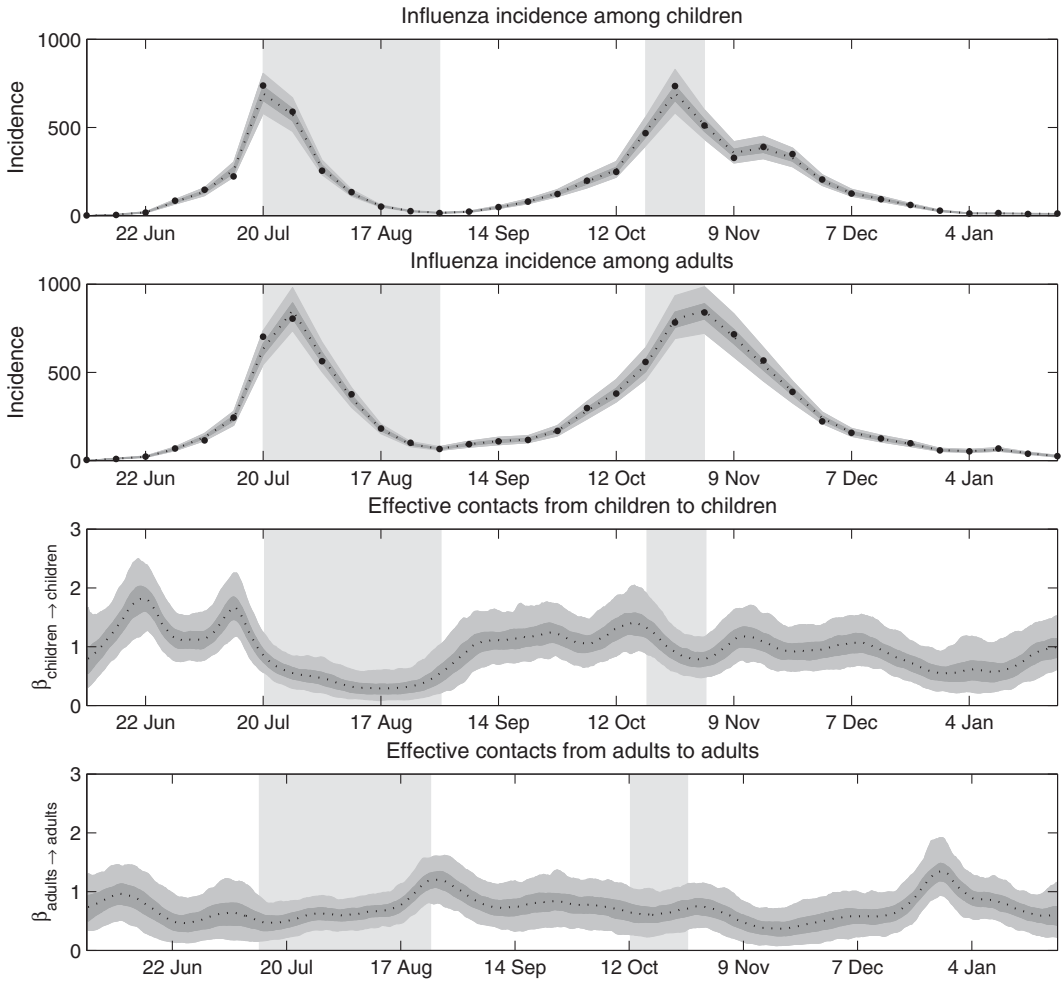

Fig. 5. Offline estimates of the effective contact rate among children and adults during the A/H1N1 2009 influenza pandemic using a 2-classes age-structured model and age-specific incidence data. Black dots indicate observed incidence estimates among each age group provided by the Health Protection Agency (first and second panels). Black dotted lines indicate the mean of the point-wise posterior density. Dark and light grey areas, respectively, indicate 50% and 95% credible intervals of the posterior density. Holidays are indicated by a very light grey area. First panel: HPA estimates of the weekly total number of A/H1N1 influenza cases among children in London (per 100 000 inhabs.) Second panel: HPA estimates of the weekly total number of A/H1N1 influenza cases among adults in London (per 100 000 inhabs.). Third panel: offline estimates of the effective contact rate from children to children. Fourth panel: offline estimates of the effective contact rate from adults to adults.

be similar with that of Figure 2 thus stressing their role to the evolution of the epidemic. More details, including posterior summaries for the parameters and information about the priors can be found in the supplementary materials (Appendix C available at *Biostatistics* online).

6. DISCUSSION

In this paper, we examined epidemic models where some of the parameters are represented by diffusions or integrals thereof. The main motivation was to account for various time-varying drivers (virus evolution,

seasonality, schools closure, etc.), while maintaining a simple interpretation. We present a unified framework that supports data augmentation MCMC schemes based on fine partitions on the diffusion path. The associated approximation error can be controlled by the user without affecting the MCMC performance and can be viewed as an extension of the approaches by Roberts and Stramer (2001); Chib *and others* (2006) to the more challenging observation regime of this paper. The consideration of the algorithms in a continuous time setting revealed major issues associated with Gibbs data-augmentation schemes. This justifies the use of particle MCMC, which updates paths and parameters jointly, while pointing directions for future research on Gibbs schemes. We also presented a computational machinery based on the PMCMC algorithm (Andrieu *and others*, 2010), which was integrated in an adaptive MCMC context. We consider EKF-based adaptive algorithms that can offer substantial improvement, especially in cases with many static parameters. This paper is one of the first applications of PMCMC in epidemic models and data; standard PMCMC schemes were also used in Rasmussen *and others* (2011).

Initially we relied on a simple SEIR model but such an analysis can be viewed as an exploratory tool towards more structured models; e.g. the age-structured model of Section 5.3 that appears to be an improved representation of reality. This approach can help in developing richer models and testing alternative scenarios for public health interventions, or to bring further insights on extrinsic factors such as climate on the dynamics of epidemics. Moreover, this framework can support multiple sources of data, of potentially different nature: Rasmussen *and others* (2011) has shown how time series and genealogies can be combined in a PMCMC inference framework for more informative estimates. While we worked mainly with influenza time series, the developed methodology can be applied to other cases; current work considers its application as part of the CHARME project (Boily *and others*, 2007). The presented approach may also be thought of as an alternative to the white noise modeling of environmental stochasticity introduced in Bretó *and others* (2009), as it offers the possibility to capture the dynamics of environmental drivers. A potential next step will be to combine environmental with demographic stochasticity, modeling infections as Poisson processes which rates depend on a time-varying β.

The inferential framework presented in this article shares the "plug and play" feature of the Iterated Filtering methodology. While extra care and further study is required for specific models or datasets, its algorithmic aspects can be decoupled from the modeling aspects. This provides the possibility to develop generic inference packages: we are currently working towards its integration in a generic inference platform inspired from the R package POMP.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors thank John Edmunds, Nikolaos Demiris, Wicher Bergsma, and the three anonymous reviewers for their helpful and constructive comments. *Conflict of Interest:* None declared.

Funding

J.D. has been supported by the Statistics department of the London School of Economics.

References

Anderson, R. M. and May, R. M. (1992). *Infectious Diseases of Humans: Dynamics and Control*. New York: Oxford University Press.

ANDRIEU, C., DOUCET, A. AND HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society: Series B* **72**, 269–342.

BAGUELIN, M., HOEK, A. J. V., JIT, M., FLASCHE, S., WHITE, P. J. AND EDMUNDS, W. J. (2010). Vaccination against pandemic influenza A/H1N1 in England: a real-time economic evaluation. *Vaccine* **28**, 2370–84.

BESKOS, A., PAPASPILIOPOULOS, O., ROBERTS, G. O. AND FEARNHEAD, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 333–382.

BOILY, M. C., LOWNDES, C. M., VICKERMAN, P., KUMARANAYAKE, L. *and others* (2007). Evaluating large-scale HIV prevention interventions: study design for an integrated mathematical modelling approach. *Sexually Transmitted Infections* **83**, 582.

BRETÓ, C., HE, D., IONIDES, E. AND KING, A. (2009). Time series analysis via mechanistic models. *The Annals of Applied Statistics* **3**, 319–348.

CAUCHEMEZ, S., BOËLLE, P. Y., THOMAS, G. AND VALLERON, A. J. (2006). Estimating in real time the efficacy of measures to control emerging communicable diseases. *American journal of Epidemiology* **164**, 591–597.

CAUCHEMEZ, S. AND FERGUSON, N. M. (2008). Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in london. *Journal of the Royal Society Interface* **5**, 885–897.

CAUCHEMEZ, S., VALLERON., A. J., BOËLLE, P. Y., FLAHAULT, A. AND FERGUSON, N. M. (2008). Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature* **452**, 750–754.

CAZELLES, B. AND CHAU, N. P. (1997). Using the Kalman filter and dynamic models to assess the changing HIV/AIDS epidemic. *Mathematical Biosciences* **140**, 131–154.

CAZELLES, B., CHAVEZ, M., MCMICHAEL, A. J. AND HALES, S. (2005). Nonstationary influence of Ei Nino on the synchronous dengue epidemics in Thailand. *PLoS Medicine* **2**, 313.

CHIB, S., PITT, M. K. AND SHEPHARD, N. (2006). Likelihood based inference for diffusion driven state space models. Working paper. http://apps.olin.wustl.edu/faculty/chib/techrep/sde.pdf

CHOPIN, N., JACOB, P. E. AND PAPASPILIOPOULOS, O. (2011). SMC^2: A sequential Monte Carlo algorithm with particle Markov chain Monte Carlo updates. arXiv:1101.1528v3.

CORI, A., BOËLLE, P. Y., THOMAS, G., LEUNG, G. M. AND VALLERON, A. J. (2009). Temporal variability and social heterogeneity in disease transmission: the case of SARS in Hong Kong. *PLoS Computational Biology* **5**, e1000471.

DOUCET, A. AND JOHANSEN, A. M. (2009). A tutorial on particle filtering and smoothing: fifteen years later. In: Crisan, D. and Rozovskii (editors), *The Oxford Handbook of Nonlinear Filtering*. Oxford: Oxford University Press, pp. 656–704.

DUKIC, V. M., LOPES, H. F. AND POLSON, N. (2009). Tracking flu epidemics using Google flu trends and particle learning. Working paper. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1513705

FERGUSON, N. (2007). Capturing human behaviour. *Nature* **446**, 733–733.

FINE, P. E. M. AND CLARKSON, J. A. (1982). Measles in England and Wales: an analysis of factors underlying seasonal patterns. *International Journal of Epidemiology* **11**, 5.

GEYER, C. J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science* **7**, 473–483.

GOLIGHTLY, A. AND WILKINSON, D. (2008). Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis* **52**, 1674–1693.

GRIFFIN, J. T., GARSKE, T., GHANI, A. C. AND CLARKE, P. S. (2011). Joint estimation of the basic reproduction number and generation time parameters for infectious disease outbreaks. *Biostatistics* **12**, 303.

HE, D. H., DUSHOFF, J., DAY, T., MA., J. AND EARN, D. J. D. (2011). Mechanistic modelling of the three waves of the 1918 influenza pandemic. *Theoretical Ecology* **4**, 283–8.

Ionides, E. L., Bretó, C. and King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* **103**, 18438.

Kalogeropoulos, K. (2007). Likelihood-based inference for a class of multivariate diffusions with unobserved paths. *Journal of Statistical Planning and Inference* **137**, 3092–3102.

Kurtz, T. G. (1981). *Approximation of Population Processes*. Philadelphia: Society for Industrial Mathematics.

Miller, E., Hoschler, K., Hardelid, P., Stanford, E. *and others*. (2010). Incidence of 2009 pandemic influenza a H1N1 infection in England: a cross-sectional serological study. *Lancet* **375**, 1100–1108.

Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications*. Berlin: Springer.

Pokern, Y., Stuart, A. M. and Wiberg, P. (2009). Parameter estimation for partially observed hypoelliptic diffusions. *Journal of the Royal Statistical Society: Series B* **71**, 49.

Rasmussen, D. A., Ratmann, O. and Koelle, K. (2011). Inference for Nonlinear epidemiological models using genealogies and time series. *PLoS Computational Biology* **7**, e1002136.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics* **18**, 349–367.

Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings Algorithm. *Biometrika* **88**, 603–621.

Särkkä, S. and Sottinen, T. (2008). Application of Girsanov theorem to particle filtering of discretely observed continuous-time non-linear systems. *Bayesian Analysis* **3**, 555–584.

Shaman, J. and Kohn, M. (2009). Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences* **106**, 3243.

Unkel, S., Farrington, C., Garthwaite, P. H., Robertson, C. and Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A* **175**, 49–82.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: Society for Industrial Mathematics.

Whitaker, H. J. and Farrington, C. P. (2004). Infections with varying contact rates: application to varicella. *Biometrics* **60**, 615–623.

[*Received February 1, 2012; revised November 2, 2012; accepted for publication November 16, 2012*]