Towards Bayesian Parametrization of national scale epidemics

Eriksson Robin * **Engblom Stefan** * **Widgren Stefan** *,**

* *Department of Information Technology, Uppsala University, 751 05 Uppsala, Sweden, (e-mail: robin.eriksson@it.uu.se, stefane@it.uu.se, stefan.widgren@it.uu.se)*
** *Department of Disease Control and Epidemiology, National Veterinary Institute, 751 89 Uppsala, Sweden, (e-mail: stefan.widgren@sva.se)*

Keywords: Simulation of stochastic systems, Discrete event modeling and simulation, Spatial stochastic models, Bayesian parameter estimation.

1. INTRODUCTION

Many infections or diseases that pose a public health threat have a *zoonotic* origin, i.e., are transmitted from animals to humans by contact with infected animals. Verotoxin-producing *Escherichia coli* (VTEC) is an example of a zoonotic foodborne pathogen where cattle can act as a reservoir, see Newell et al. (2010). Livestock movements are the primary transmission route for transferring VTEC infections between cattle herds, Nielsen et al. (2002). EU regulations require member states to keep national databases of all bovine animals and it is therefore possible to develop realistic large-scale disease spread models that incorporate the transport network to better understand the transmission of zoonotic infections in the cattle population. Although mainly inspired by zoonotic diseases and models driven by livestock data, our discussion is of entirely general character and applies to arbitrary epidemiological models. We have implemented a framework for stochastic disease spread simulator on networks in the software **SimInf**, see Widgren et al. (2016a), which is a `C` compiled extension to the programming language `R` available through the Comprehensive R Archive Network (CRAN). With **SimInf** and data with detailed information about the movement of the Swedish cattle population and bacterial testing at multiple sites we can perform Bayesian parameter inference on national scale epidemics.

2. EPIDEMIOLOGICAL MODELING

The SIS_E-model consists of the two compartments susceptible (`S`) and infected (`I`) and an environmental compartment (`E`) representing an infectious pressure from free-living pathogens. The infection transmits indirectly from infected to susceptible individuals through the local environment, contaminated by infected individuals. Within each herd i, the SIS_E model has the following two state transitions,

$$\left.\begin{array}{l} S_i \xrightarrow{\upsilon\varphi_i} I_i \\ I_i \xrightarrow{\gamma} S_i \end{array}\right\}, \tag{1}$$

where υ is the indirect transmission rate of the environmental infectious pressure, and γ is the recovery rate from the infection. Moreover, $\varphi_i(t)$ is the concentration of the local environmental-infectious pressure in herd i, evolved as

$$\frac{d\varphi_i(t)}{dt} = \frac{\alpha I_i(t)}{S_i(t) + I_i(t)} - \beta(t)\varphi_i(t), \tag{2}$$

where α is the average shedding rate of bacteria to the environment per infected individual, while the time-dependent function β captures the decay and removal of bacteria. The model can be extended to include multiple compartments, such as different age groups in the susceptible and infected compartments, see Bauer et al. (2016); Widgren et al. (2016b). With the inclusion of observations, we implement (1) and (2) as stochastic simulations on a connected network in **SimInf**, see Engblom and Widgren (2017); Bauer et al. (2016).

The data we have available contains a total of 18,649,921 reports with information about; first, the date and the node for birth events, second, the date, the source, and destination node for any movements, and third, the date for slaughter or death, Nöremark et al. (2011). Each unique node identifier (n = 37,221) in the data corresponds to a single geographical location where animals are kept, and could, e.g., correspond to a farm building or pasture distributed across the entire Sweden.

3. BAYESIAN PARAMETRIZATION

We consider a postulated truth in the form of a time-dependent stochastic process $X(t) = X(t, \theta)$, for some parameter θ. The density for this process is denoted by $P\big((x,t)|(x',t');\theta\big) = \mathbf{P}\big(X(t) = x|X(t') = x';\theta\big)$, and we consider throughout this work that the true density is computationally intractable but — mainly for convenience — is Markovian. We are given a set of observations $(x_i) = (x_i, t_i) \sim X(t_i)$, and the task is to estimate the unknown parameter θ.

In exploring the posterior density $\mathbf{P}(\theta|x) \propto \mathbf{P}(x|\theta)$, likelihood-based inference methods are not viable and other methods need to be advised. We consider two likelihood-free inference methods, first, Approximate Bayesian Computations (ABC), see Beaumont et al. (2002), and second, Synthetic Likelihood Markov chain

Monte Carlo (SLMCMC) as in Wood (2010). Both methods generate simulated data $(z_i(\theta')) \sim X(t_i, \theta')$ and compare it with the observed data as a substitute for the likelihood.

In ABC one compares the summarized versions of $\{(z_i), (x_i)\}$, the summary statistics $\{S(z), S(x)\}$, using a distance measure, e.g., the Euclidean norm. If the distance is smaller than a tolerance ϵ, the proposed parameter θ' is accepted. The ABC method thus gives the approximate posterior distribution defined as

$$\mathbf{P}_\epsilon(\theta|S(x)) \propto \int_{\mathcal{X}} \mathbf{P}(z|\theta')\mathbf{P}(\theta')\mathbb{I}_{A_{\epsilon,x}}(z)\mathrm{d}z \qquad (3)$$
$$A_{\epsilon,x}(z) = \{z \in \mathcal{X}; ||S(z) - S(x)|| < \epsilon\}.$$

The choice of the acceptance tolerance ϵ will define how close to the true posterior the approximation is. As $\epsilon \to \infty$ the sample distribution is the prior: $\mathbf{P}_\epsilon(\theta|S(x)) \to \mathbf{P}(\theta)$, and as $\epsilon \to 0$ the approximation will converge to the posterior $\mathbf{P}_\epsilon(\theta|S(x)) \to \mathbf{P}(\theta|S(x))$, see Wilkinson (2013).

The other method referred to as SLMCMC considers each set of simulated summary statistics to be an observation of a multivariate normal distribution $S(\cdot) = \boldsymbol{s} \sim \mathcal{N}(\boldsymbol{m}_\theta, \Sigma_\theta)$, where $\boldsymbol{m}_\theta$ is the mean and Σ_θ is the covariance. When assuming normality, we utilize an auxiliary model $\mathcal{Z}$ and will, in turn, be able to observe the auxiliary model's posterior density $\mathbf{P}_{\mathcal{Z},\eta}(\theta|\boldsymbol{s})$. The accuracy of the observed posterior density depends on the number of observation η of $\boldsymbol{s}$ and the validity of the assumption of the auxiliary model $\mathcal{Z}$ being descriptive of the postulated truth. We construct the synthetic log-likelihood as

$$\boldsymbol{p}_{\mathcal{Z},\eta}(\boldsymbol{s}|\theta) = -\frac{1}{2}(\boldsymbol{s} - \hat{\boldsymbol{m}}_\theta)^\top \hat{\Sigma}_\theta^{-1}(\boldsymbol{s} - \hat{\boldsymbol{m}}_\theta) - \frac{1}{2}\log|\hat{\Sigma}_\theta|, \quad (4)$$

where $\hat{\boldsymbol{m}}_\theta$ and $\hat{\Sigma}_\theta$ are estimates of the mean and covariance. We then explore the approximate posterior density using (4) in a likelihood-based Markov chain Monte Carlo method.

In Figure 1, we illustrate a proof of concept for the two methods. We conduct the parameter inference on a Geometric Brownian Motion, for which the likelihood function is known, and we present the results from ABC and SLMCMC together with the likelihood-based Metropolis-Hastings algorithm, Hastings (1970), i.e., the best attainable posterior in this setting. We are currently applying these likelihood-free methods to the SIS_E-model using series of measurements

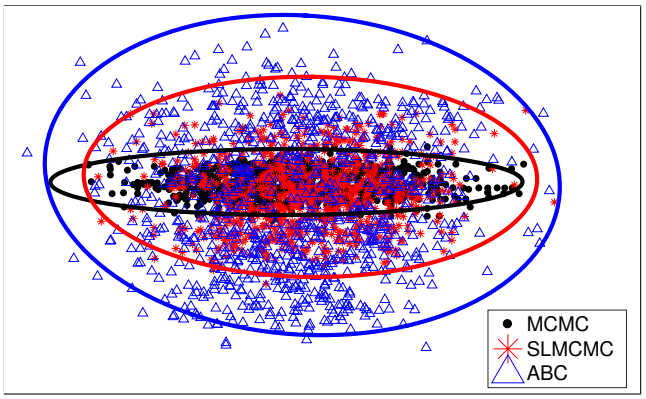

Fig. 1. Proof of concept for the likelihood-free methods, ABC and SLMCMC. Estimating parameters for a geometric Brownian motion together with a Metropolis-Hastings generated estimation present.

ACKNOWLEDGEMENTS

This work was financially supported by the Swedish Research Council Formas (S. Engblom, S. Widgren), by the Swedish Research Council within the UPMARC Linnaeus center of Excellence (S. Engblom, R. Eriksson), and by the Swedish strategic research program eSSENCE (S. Widgren).

REFERENCES

Bauer, P., Engblom, S., and Widgren, S. (2016). Fast event-based epidemiological simulations on national scales. *Int. J. High Perf. Comput. Appl.*, 30(4), 438–453. doi:10.1177/1094342016635723.

Beaumont, M.A., Zhang, W., and Balding, D.J. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4), 2025–2035.

Engblom, S. and Widgren, S. (2017). Chapter 11 - data-driven computational disease spread modeling: From measurement to parametrization and control. In A.S.S. Rao, S. Pyne, and C. Rao (eds.), *Disease Modelling and Public Health, Part A*, volume 36 of *Handbook of Statistics*, 305 – 328. Elsevier.

Hastings, W.K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1), 97–109.

Newell, D.G., Koopmans, M., Verhoef, L., Duizer, E., Aidara-Kane, A., Sprong, H., Opsteegh, M., Langelaar, M., Threfall, J., Scheutz, F., et al. (2010). Food-borne diseases - the challenges of 20 years ago still persist while new ones continue to emerge. *International journal of food microbiology*, 139, S3–S15. doi: 10.1016/j.ijfoodmicro.2010.01.021.

Nielsen, E.M., Tegtmeier, C., Andersen, H.J., Grønbaek, C., and Andersen, J.S. (2002). Influence of age, sex and herd characteristics on the occurrence of verocytotoxin-producing *Escherichia coli* O157 in danish dairy farms. *Veterinary Microbiology*, 88(3), 245–257.

Nöremark, M., Håkansson, N., Lewerin, S.S., Lindberg, A., and Jonsson, A. (2011). Network analysis of cattle and pig movements in sweden: measures relevant for disease control and risk based surveillance. *Preventive veterinary medicine*, 99(2), 78–90.

Widgren, S., Bauer, P., and Engblom, S. (2016a). Siminf: An r package for data-driven stochastic disease spread simulations. *arXiv preprint arXiv:1605.01421*.

Widgren, S., Engblom, S., Bauer, P., Frössling, J., Emanuelson, U., and Lindberg, A. (2016b). Data-driven network modelling of disease transmission using complete population movement data: spread of VTEC O157 in Swedish cattle. *Veterinary Research*, 47(1), 81. doi: 10.1186/s13567-016-0366-5.

Wilkinson, R.D. (2013). Approximate bayesian computation (abc) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, 12(2), 129–141.

Wood, S.N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310), 1102–1104.