MSc thesis proposals:
Computational Epidemics driven by Data

Stefan Engblom

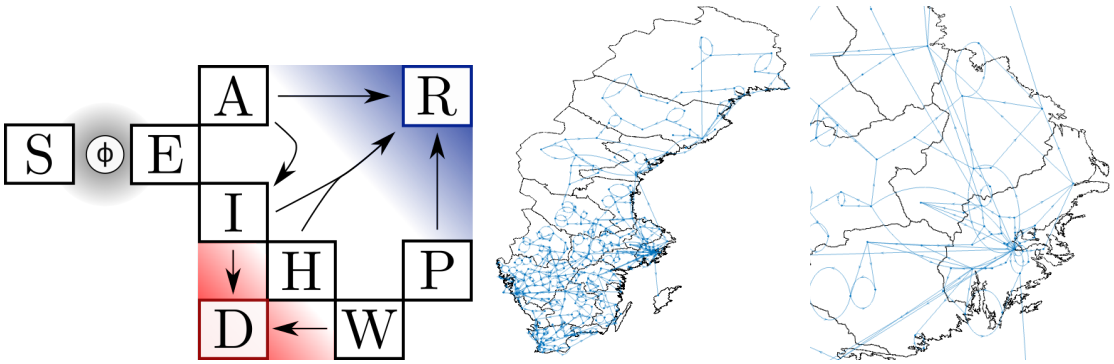

Fig. 1. *Left:* sample compartment model for Covid-19, *right:* transport network between Sweden's municipalities derived from commuting data and using nonlinear optimization techniques.

Summary

Public health agencies are increasingly dependent on various sets of data in order to accurately plan for disease scenarios, to design mitigation- and suppression programs, and to assess epidemiological risks. The quality of such strategies ultimately depends on the development of a range of tools. In these proposed projects we will work in a Matlab-based simulation software `www.urdme.org` and are particularly interested in pursuing the following five directions of research:

A. Probabilistic Epidemics: the objective of this part is to design epidemiological models which carry a likelihood. Relevant background for a candidate include computational science, sequential Monte Carlo methods, automatic control/filtering.

B. Data-driven modeling: in this project the overall goal is to determine best practice for integration of epidemiological data in a few selected concrete models. A background revolving around statistical modeling and probabilistic machine learning will be useful.

C. Bayesian inference: development of simulation-based inference and posterior exploration algorithms. Courses in computational science, optimization, and sequential Monte Carlo methods/filtering will be beneficial.

D. Model errors and data: ideally, the errors in a derived model will balance those that come from the data available to parameterize it. In this project the target is to quantify the *practical observability* in a few concrete selected models. Tools from statistical modeling and filtering will be employed in this project.

E. Model validation: given a generative model, data, and a proposed parameterization one likes to quantify how reasonable the model is in view of data. We will mainly rely on statistical modeling here but methodology from probabilistic machine learning will also be employed.

Specifics of these projects

A. Probabilistic Epidemics

A few typical epidemiological compartment models are displayed in Fig. 2. The objective of this project is to devise a simulation framework for this kind of epidemiological models *which by design carries a likelihood*. Techniques to achieve this will involve Kalman filters as well as SMC-type particle filters. For example, it is of interest to investigate the trade-offs in using a full simulator as the basis for a particle filter versus relying on approximations built from Gaussian mixture models.

Another concrete implementation issue is to support a spatial or demographic resolution via replication of models as indicated in Fig. 2 (*middle*), and equipping such a framework with a computable approximative likelihood.

The project will benefit from a working startling implementation in Matlab.

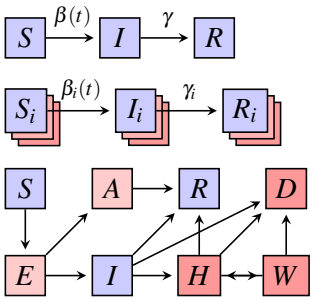

Fig. 2. *Top:* SIR model with dynamic infectivity $\beta(t)$. *Middle:* a more detailed resolution (across a network or demographic categories) by replicating the states. *Bottom:* a more involved model for Covid-19 with exposed/asymptomatic states as well as states for hospitalized and deaths.

B. Data-driven modeling

Ideally, epidemiological models are integrated with data. Relevant examples include transport- or contact networks as in Fig. 1, statistics for hospital duration/treatment success, and hazard estimates depending on, for example, age or socioeconomic factors. Another related issue is to integrate weakly informative streams of pathogen detection data consisting of 0's and 1's. A challenge to overcome here is that many of these problems involve distributions which are technically involved to handle, e.g., non-exponential waiting times or highly skewed and non-Gaussian distributions for pathogen samples.

There are several challenges to approach which are possible for this project. For *networks*, methods based on optimization for graphs will be tried out and evaluated. An example is found in Fig. 1 which was obtained using public commute data and a nonlinear fitting procedure. An idea to model non-exponential waiting times is via ordinary exponential waiting times but using varying waiting time parameters and distributed across hidden states. Finally, non-Gaussian pathogen samples can be modeled using either Gaussian mixtures or via so-called *unscented* Kalman filters.

C. Bayesian inference

There is an interesting tradeoff between *information reduction approaches* based on synthetic likelihood and MCMC, common in applied statistics, and *state estimation methods* prevalent in the automatic control community. The former tend to be more robust with respect to model misspecification, whereas the latter offer sharper inference when successfully applied.

The plan of attack in this projects is to use a Gaussian ensemble filter as a core likelihood, operating within an adaptive Metropolis scheme which manages the static parameters of the model. Some research and a bit of experimenting is needed to enable control of the regularity of the filter's estimator for the time-dependent parameters. Loosely speaking, the pathwise regularity of the dynamic parameters will be controlled by the outer Metropolis sampler for the static parameters. An example from a startling implementation in Matlab is shown in Fig. 3.

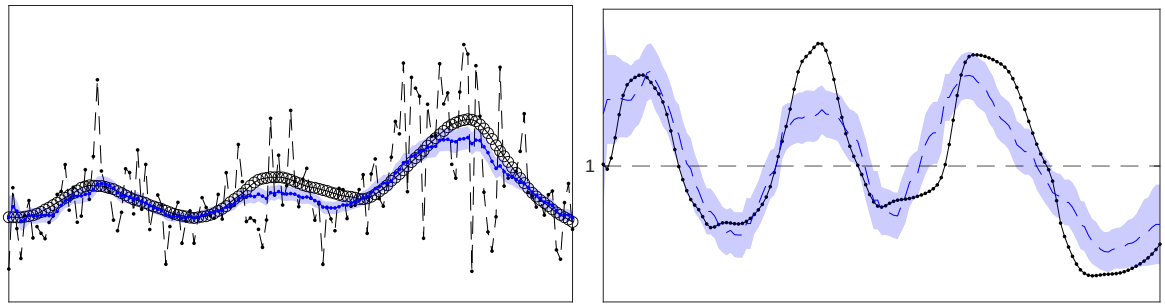

Fig. 3. *Left:* noisy measurements of virus concentration in wastewater, *right:* dynamic estimates for the *reproduction number* R_t with the known/synthetic truth in black. Blue shade is the estimated uncertainty.

D. Model errors and data

A mathematical tool to measure the identifiability of an inference problem is the *Fischer information matrix* (FIM). In this project we are more specifically interested in the *path-wise* FIM, or pFIM for short. Chiefly, this matrix measures the sensitivity of observables to changes in parameters along a selected family of simulated paths. To be useful in epidemics, it needs to be generalized to problems with time-varying parameters and also combined with the state estimation concept of *observability*. We will tackle both issues in the setting of Gaussian ensemble filters

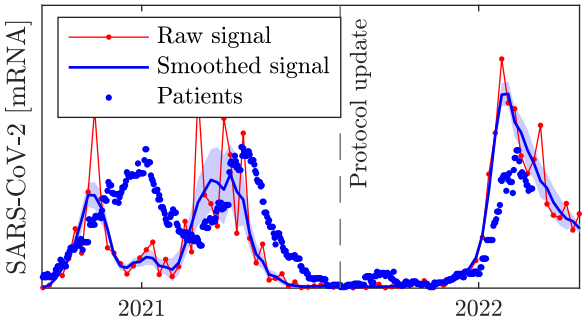

Fig. 4. SARS-CoV-2 mRNA concentration in wastewater and patients hospitalized due to Covid-19 in Uppsala county. The signal improved after the update of the experimental protocol.

The use of this is that, *before* incorporating data, say for example the weekly concentration of mRNA in wastewater as in Fig. 4, the sensitivity to parameters and the observability of states can be estimated along considered paths. *After* real data has been integrated, any posterior model can be similarly assessed. In this way, situations with weakly informative data and poor identifiability can be detected and handled.

Also for this project suggestion we have access to a few relevant data sets and a working albeit basic implementation in Matlab.

E. Model validation

The most open-ended project proposal is concerned with estimating, in view of the data, if a calibrated posterior model can be accepted/not be rejected. In this project we will attempt an approach in the form of a *data consistency check* (DCC) for models in a class, i.e., sampled from a set of competing

models. The approach is attractive since it is highly interpretable and transparent, but the downside is that the procedure as a whole is computationally very demanding.

The DCC is fast enough to conveniently employ for smaller epidemiological sub-models where Monte Carlo sampling is efficient. Concrete examples to be investigated and for which we already have access to data include (1) non-Markovian waiting times for hospital duration, (2) models for delayed reporting, and (3) wastewater signal as a function of the infectious population.

Contact

Candidates with a background in one or more of Scientific Computing, Automatic Control, Applied Mathematics, Data Science are more than welcome to contact me for more information.

- Mail: stefane@it.uu.se
- URL: https://stefanengblom.github.io/