UPPSALA
UNIVERSITET

Bayesian epidemiological modeling: with little and without data

Stefan Engblom

Div of Scientific Computing, Dept of Information Technology, Uppsala University, Uppsala, Sweden

EMMC-eSSENCE 2019, June 3–6 2019, Uppsala

Outline

Bayesian epidemics

1. With little data

2. Without data

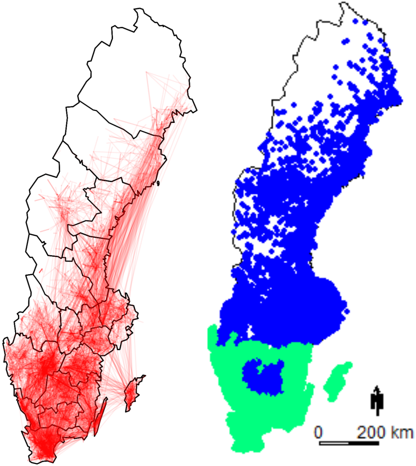
3. Conclusions

$\Longrightarrow$ Joint work with **Robin Eriksson** @ Dept of IT, Uppsala university, and **Stefan Widgren** @ Dept of Disease Control and Epidemiology, National Veterinary Institute (SVA). $\Longleftarrow$

Case study: modeling the spread of VTEC O157

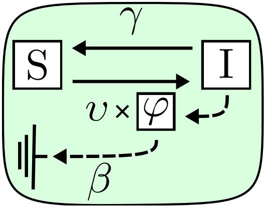
Verotoxinogenic *E. coli* O157:H7 in the Swedish cattle population

- *Zoonotic pathogen* (animal $\rightarrow$ human) of great public health interest
- Substantial amount of data:
  - individual-level cattle data from 2005 and onwards
  - meteorological data
- Less data:
  - actual disease measurements at farms (enough for parametrization?)

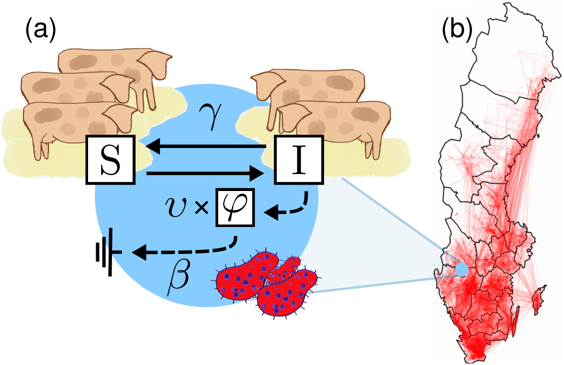


The SIS$_E$ model

Replicated across a data-driven network



Susceptible individuals, Infected individuals, and φ, the infectious pressure.

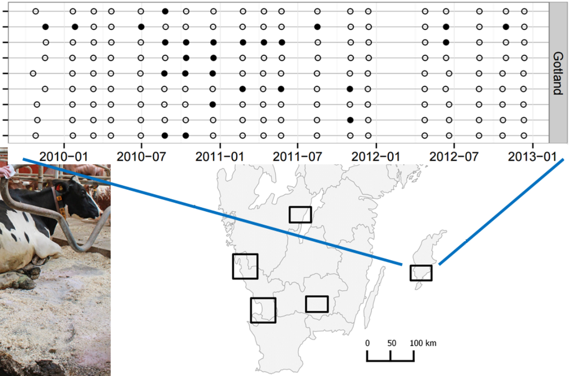


SimInf

www.siminf.org

Severely limited by data

126 out of 37,221 holdings were sampled once every 6 to 8 weeks for 38 months; so disease data is 6–8 binary true/false samples per year at 0.3% of the nodes. Also, the sensitivity of the test had to be estimated...



Synthetic Likelihood Adaptive Metropolis ("SLAM")

Bayesian computations with untractable likelihoods

- Multiple simulations z_i for a proposed θ; $z_\theta = (z_1, z_2, \ldots, z_N)$.
- Assume that some *summary statistics* $S(\cdot)$ is an observation from a multivariate Gaussian distribution $\mathcal{N}(\mu_\theta, \Sigma_\theta)$, estimated by

$$\hat{\mu}_\theta = \frac{1}{N}\sum_{i=1}^{N} S(z_i)$$

$$\widehat{\Sigma}_\theta = \frac{1}{N-1}(\mathbf{S} - \hat{\mu}\mathbb{1}^{(N)})(\mathbf{S} - \hat{\mu}\mathbb{1}^{(N)})^\top$$

- We get the "synthetic" likelihood $P(s_{\text{obs}}|\mathbf{S}) = \mathcal{N}(s_{\text{obs}}|\hat{\mu}_\theta, \widehat{\Sigma}_\theta)$

SLAM sampling:

Consider initial $(\theta^{(1)}, \mathcal{L}_\theta)$ and summarized data s_{obs}.
for $i = 2, \ldots, N_{\text{sample}}$ **do**
  Compute $C^{(i)} = \xi_d \operatorname{Cov}(\theta^{(1)}, \ldots, \theta^{(i-1)}) + \xi_d \epsilon I_d$
  Propose $\theta^* \sim \mathcal{N}(\theta^{(i-1)}, C^{(i)})$
  Simulate $Y = (y_1, \ldots, y_N), y_j \sim F(\theta^*)$
  Bootstrap $Z = (z_1, \ldots, z_R), z_j \sim \hat{F}_N(Y)$
  Estimate $(\hat{\mu}_{\theta*}, \widehat{\Sigma}_{\theta*})$ from $S = \mathbf{S}(Z)$
  Compute $\mathcal{L}_{\theta*} = P(s_{\text{obs}}|\mathbf{S})$
  if $\mathcal{U}(0,1) < \min\left(1, \mathcal{L}_{\theta*}/\mathcal{L}_\theta\right)$
    $\theta^{(i)} = \theta^*$ and $\mathcal{L}_\theta = \mathcal{L}_{\theta*}$
  else
    $\theta^{(i)} = \theta^{(i-1)}$

A series of inverse crimes

Navigating through a forest of complexity

Basic idea: Solve a series of increasingly realistic inverse problems using known truth data until the desired set-up is reached.

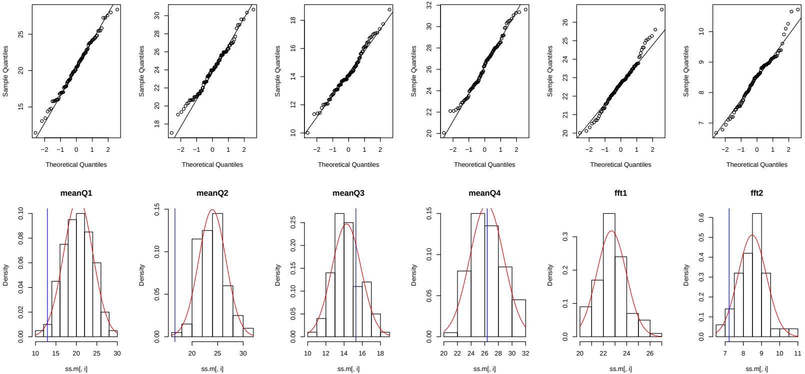
Personal reflections

- Model correctness cannot be assumed
- Identifiability cannot be assumed
- Real data is *much worse* than synthetic data
- The main *insight* comes from solving problems on the way

Suitable summary statistics?

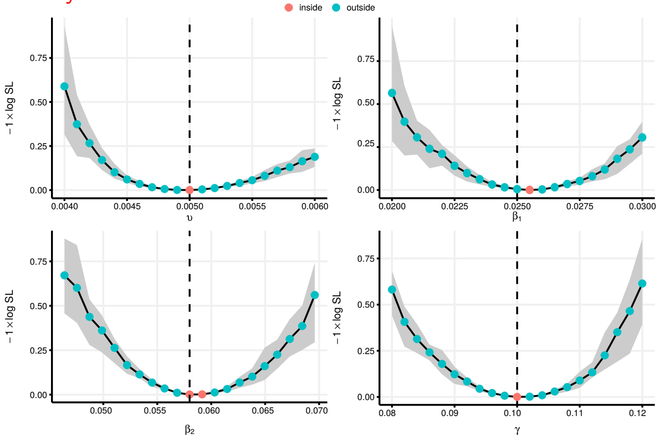
N parameters $\longrightarrow$ find at least N SS.

- Need "normal"-like SS for the SL ansatz
- And they are!



Feasible optimization?

- Need that the $(-\log \mathrm{SL})$ minima are well defined in each parameter dimension
- And they are!



Finally, full model results

$\sim$5% std error

Real network & actual observations

- From the mean posterior estimate, $\hat{\theta}$, we construct new synthetic data and bootstrap to estimate the bias
- Posterior use: evaluate surveillance- and mitigation strategies probabilistically

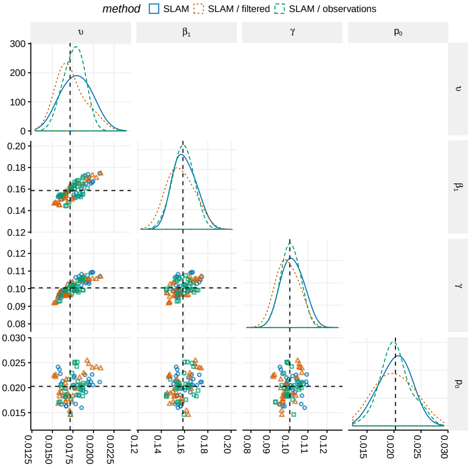


Figure: Posterior samples.

Case study: spread of Antimicrobial Resistance (AMR)

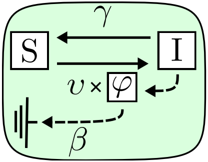
Question-driven rather than data-driven modeling

How can we understand the 'flow' of AMR spread?
- "understand" $\sim$ identify the dominating processes and their timescales, estimate qualitatively, or simply get a feeling for...

BUT: No "hard" data to easily build models on!

The SIS_E framework again

Being verotoxinogenic is caused by a certain strand, and so is resistance to antibiotics:

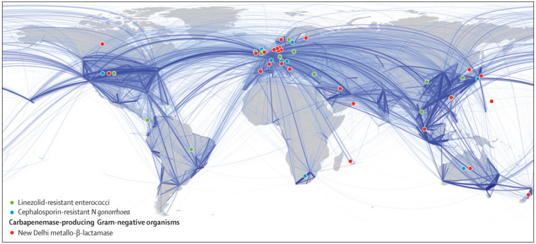


1. $\{\gamma, \beta\}$ set the time scale of recovery and open space decay of bacteria, respectively.
2. Hence υ alone determines the stationary prevalence.

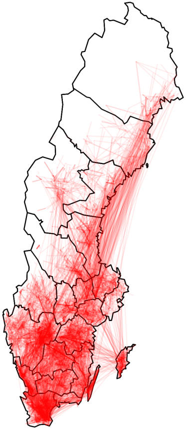
So, *the latent variables* (AMR fitness & antibiotic pressure) $\xrightarrow{\text{proxy}} \upsilon \xrightarrow{t \to \infty} P_\infty$, the stationary prevalence.

Network data

Sample real networks



(a) Worldwide travel routes and emergence of antimicrobial resistance Source: Holmes *et al.*, "Understanding the mechanisms and drivers of antimicrobial resistance", Lancet 387 (2016)



(b) Cattle network data: ∼10 years of data, ∼40,000 nodes

Model reduction

Bayesian homogenization

Ansatz borrowed from statistical physics: SDE in *gradient form* for the prevalence $P(t) := I(t)/N(t) \in [0, 1]$,

$$dP(t) = -V'(P)\,dt + \sigma\,dW(t),$$

where V is the *epidemic potential energy*.

-We can find V and σ by many full simulations over a range of the (proxy) parameter using (Variational-) Bayes techniques.

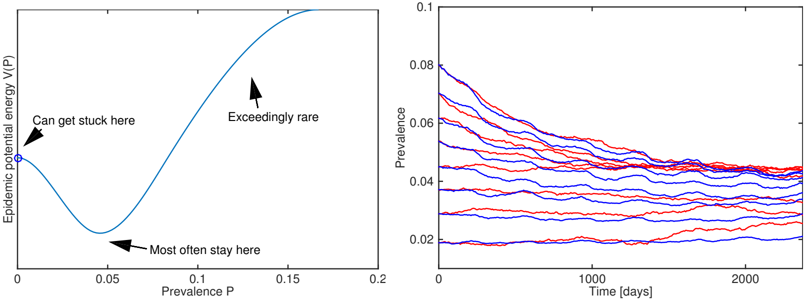
-Fokker-Planck equation for density $\rho(t, P)$, known stationary (Gibbs) distribution:

$$\rho_t = [V'(P)\rho]'_P + \frac{\sigma^2}{2}[\rho]''_{PP} \qquad + \text{certain BCs},$$
$$\rho_\infty(P) \propto \exp(-2\sigma^{-2}V(P)).$$

The SDE form fascilitates detailed computational analysis.

Homogenized SDE

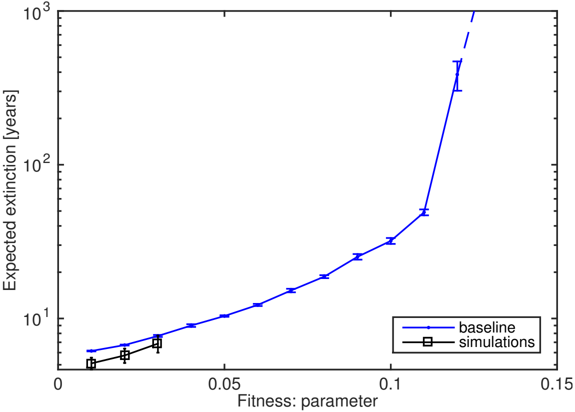
What it looks like



Left: epidemic potential $V(P)$, *right:* (blue) data from full model, (red) homogenized SDE model.

Endemic or not?

Courtesy of the Fokker-Planck

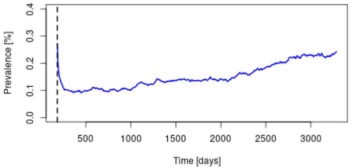


Very strong nonlinear response $\Longrightarrow$ new question: what is the effect if nodes experience a heterogeneous antibiotic pressure?

Locally increased antibiotic pressure

According to in-degree: hospitals, schools, resorts...

- The antibiotic pressure is set higher in the top-0.1% in-degree nodes
- Everywhere else the conditions are such that extinction within a few years can be expected
- *Result:* the nonlinear response makes the full system endemic for indefinite times



Conclusions

Bayesian epidemiological modeling

With little data:

1. Put effort into the model itself, this is part of the prior
2. Use inverse crimes to ensure identifiability ($\Longrightarrow$ bootstrap)
3. Synthetic Likelihood Adaptive Metropolis (SLAM) performed well

Without data:

1. Question-driven modeling $\Longrightarrow$ identify proxy variables (& proxy data)
2. Effective gradient SDE model enabled a detailed computational analysis not possible from simulations alone

Thanks for listening!