Inference with Implicit Likelihoods for Infectious Disease Models

Murali Haran

Department of Statistics, Pennsylvania State University

Applied Bayesian and Computational Statistics Working Group, University of Washington

January 2012

Collaborators

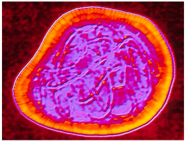
- Roman Jandarov, Department of Statistics, Penn State University
- Ottar Bjørnstad, Center for Infectious Disease Dynamics, Penn State University
- Matthew Ferrari, Center for Infectious Disease Dynamics, Penn State University
- Bryan Grenfell, Department of Ecology and Evolutionary Biology, Princeton University

Support from the Bill & Melinda Gates Foundation.

BILL & MELINDA GATES foundation

CIDD CENTER FOR INFECTIOUS DISEASE DYNAMICS

Understanding Measles Dynamics



Rubella: photo from sciencephoto.com

Photo from www.measlesrash.org

Photo from www.news.bbc.co.uk

- Measles is one of the leading causes of death among young children globally.
- An estimated 164,000 people died from measles in 2008.
- Measles is common in many developing countries.
- Goal: Understanding metapopulation dynamics and effects of spatial coupling in measles transmission. Epidemic responses.

Common Challenges

- Complex models
- The data are spatiotemporal and high dimensional
- Lots of latent variables
- Traditional likelihood-based inference is problematic:
  - Fitted models may not capture the important biological characteristics
  - May lead to poor parameter estimates
  - Computationally challenging

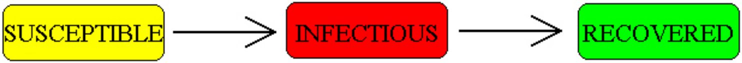
This Talk

- A new inferential approach that simultaneously addresses
  - Computational challenges
  - Inferential issues
- Motivating example: the Gravity Time series Susceptible-Infectious-Recovered (TSIR) model for measles dynamics.

SIR Model Basics

- Susceptible-Infectious-Recovered (SIR) model is an important model for infectious diseases.
- The population is subdivided into distinct classes: individuals are either susceptible (S), infectious (I) or recovered (R).
- An SIR model describes the dynamics of the sizes of each group.

Assumptions of Basic SIR Model



- Susceptible:
  - Individuals are born into this class.
  - They have never come into contact with the disease.
  - They can become infected. If infected, they move into the infectious class.
- Infectious: Individuals spread the disease to susceptibles. They remain in this class for an “infectious period” before moving into the recovered class.
- Recovered class individuals are immune for life.

Gravity TSIR Model

- Model for number of cases of measles in K cities.
- Has components of a discrete time-series SIR model (Bjørnstad et al., 2002; Grenfell et al. 2002).
- Includes seasonality in the transmission rates.
- Allows for spatial transmission between different cities.
- Models stochasticity in disease transmission and immigration.

Xia, Bjørnstad and Grenfell (2004).

Gravity TSIR Model: Notation

- Variables:
  - I_{kt} : number of infectious individuals in city k at time t
  - S_{kt} : number of susceptible individuals in city k at time t
  - L_{kt} : number of infectious people moved to city k at time t
  - d_{kj} : distance between cities k and j
  - N_{kt}, B_{kt} : size and birth rate of city k at time t
- Parameters:
  - For local dynamics: α and β (Bjørnstad et al. 2001)
  - For spatial transmission: θ, τ_1, τ_2 and ρ

Gravity TSIR Model

For city *k* at time *t*:

- # of incidences

$$I_{k(t+1)} \sim \text{Poisson}(\beta_t S_{kt}(I_{kt} + L_{kt})^{\alpha})$$

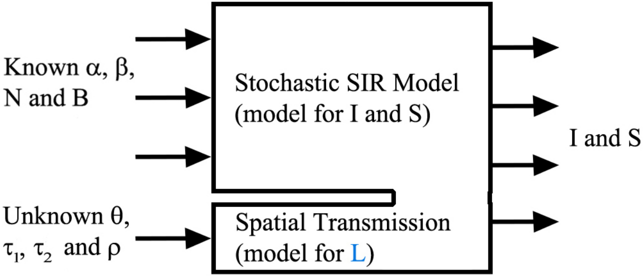
- # of susceptibles

$$S_{k(t+1)} = S_{kt} + B_{kt} - I_{k(t+1)}$$

- # of infectious immigrants (latent)

$$L_{kt} \sim \text{Gamma}\left(\theta N_{kt}^{\tau_1} \sum_{j=1, j\neq k}^{K} \frac{(Ijt)^{\tau_2}}{d_{kj}^{\rho}}, 1\right)$$

Gravity TSIR Model: Graph



Known α, β, N and B

Unknown θ, τ_1, τ_2 and ρ

Stochastic SIR Model (model for I and S)

Spatial Transmission (model for L)

I and S

I - # of cases

S - # of susceptibles

L - influx of infection

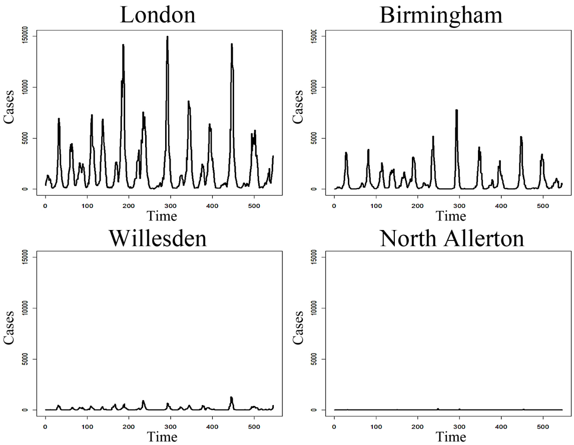
N - population sizes

B - # of births

Inference for Measles Dynamics

- Sources of information:
    - The UK Registrar General's data for 952 cities in England and Wales for years 1944-1966 of biweekly incidences of measles.
    - Number of susceptibles from standard reconstruction algorithms (cf. Fine and Clarkson 1982a, Finkenstadt and Grenfell 2000).
- **Goal**: Infer gravity parameters $\boldsymbol{\Theta} = (\theta, \tau_1, \tau_2, \rho)$ from data.

Measles Data



Notice: 952 cities of varying sizes and levels of “infecteds.”

Complicates likelihood-based inference.

Likelihood Evaluations

Why is it expensive to evaluate the likelihood?

- If $I = \{I_{kt}\}$ (infectious), $L = \{L_{kt}\}$ (latent transient infection) and $\mathbf{\Theta} = \{\theta, \tau_1, \tau_2, \rho\}$,

$$\mathcal{L}(I|\mathbf{\Theta}) = \int_L \prod_{k=1}^{K} \prod_{t=1}^{T-1} \mathcal{L}(I_{k(t+1)}|I_{kt}, L_{kt}) \times \mathcal{L}(L_{kt}|I_{kt}, \mathbf{\Theta}) dL.$$

- Requires integration over $T * K$ unobserved $\{L_{kt}\}$'s.
- Each evaluation of $\mathcal{L}(L_{kt}|I_{kt}, \mathbf{\Theta})$ for all k and t requires many summations.

Simplifications and Gridded MCMC

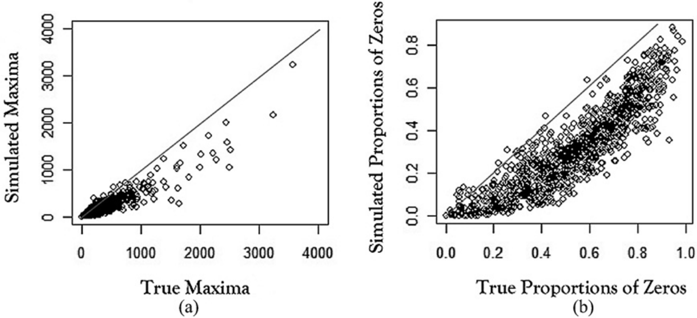
- A possible solution:
  1. We simplify by fixing the number of immigrants (latent variables) at their expected values. Likelihood function is still expensive.
  2. Discretize parameter space, pre-calculate expensive parts of the likelihood ahead of time, in parallel.
- Good news: Greatly speeds up computing, permits maximum likelihood and Bayesian inference.
- Problems . . .

Important Biological Characteristics

What do the biologists care about? “Signatures” of the process:

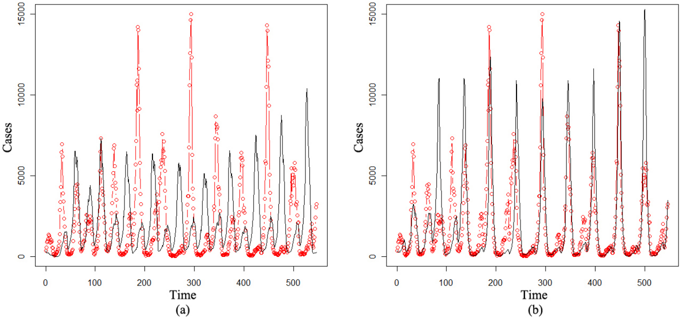
- Maximum number of incidences. $\mathbf{M} = (M_1, \cdots, M_K)$, where M_i is the maximum number of incidences for i-th city.
- Proportions of biweeks without any cases of the disease. $\mathbf{P} = (P_1, \cdots, P_K)$, where P_i is the proportion of incidence free bi-weeks for i-th city.

Problems with Fitting Key Characteristics



Fitted model does not capture well important biological characteristics of the observations.

Problems with Prediction

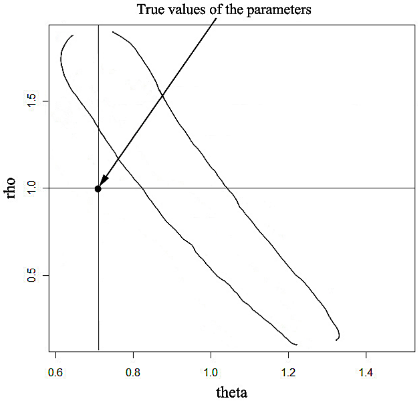


Red: observations, black: predictions

(a) likelihood-based

(b) using different (lower likelihood) parameter setting

Problems with Inference



95% confidence region for (θ, ρ)

Likelihood-based approach does not recover Θ.

Motivation for a New Approach

- Likelihood-based approaches do not give enough importance to features that are of scientific interest.
- Inference for the parameters is poor.
- Need an alternative method that:
  - Focuses on scientifically important features of the data
  - Resolves inferential issues
  - Allows for fast computations
- Cost of simulations make approximate Bayesian computation (cf. Pritchard et al., 1999) infeasible.

Gaussian Processes

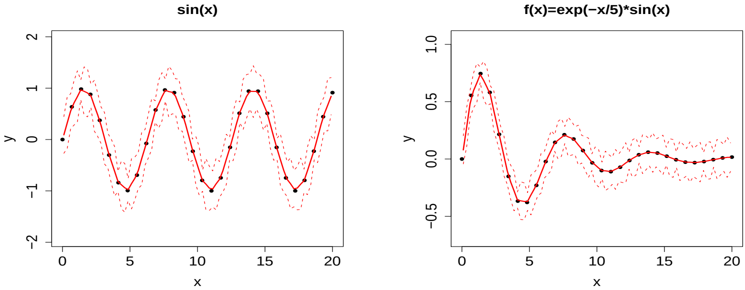
- Our new approach is based on Gaussian process emulation of the infectious disease model.
- Review: A stochastic process , $\{X(s), s \in E \subset R^d\}$, $d \geq 1$ is called a Gaussian process if for any $k > 0$ and $s_1, \cdots, s_k \in E$, $(X(s_1), \cdots, X(s_k))$ is a k-dimensional multivariate normal random variable.

Modeling with Gaussian Processes

- Gaussian processes (GPs) are useful models for:
  - Dependence e.g. time series, spatial data
  - Complicated functions

    Key idea: dependence (spatial random effects) adjusts for non-linear relationships between input and output.
- Applications:
  - Used in modeling space-time processes (cf. Cressie, 1993)
  - Emulation and calibration of complex computer models (cf. Sacks et al., 1989; Bayarri et al., 2007; Bhat et al., 2010)
  - Machine learning (cf. Rasmussen and Williams, 2005)

GP for Function Approximation: 1-D Example



The red curves are interpolations using *the same, simple GP model* with constant mean μ:

$y(x) = \mu + w(x)$, $\{w(x), x \in (0, 20)\}$ is a zero-mean GP.

An Emulation-Based Solution

- Let vector of summary statistics from observations be **Z**.
- Simulate realizations of the gravity TSIR model at p different parameter settings $\Theta_1, \Theta_2, \ldots, \Theta_p$.
- Let $\mathbf{Y}(\Theta)$ be the vector of summary statistics obtained at parameter setting Θ.
- Consider: $(\Theta_1, \mathbf{Y}(\Theta_1)), \ldots, (\Theta_p, \mathbf{Y}(\Theta_p))$.
- Stochastic emulation: fit a Gaussian process (GP) to above simulations.
  - Thus for any new parameter setting Θ^*, we have a predictive distribution for the process $\mathbf{Y}(\Theta^*)$.

Our Inferential Approach

1. Emulation: Fit a Gaussian process to $(\Theta_1, \mathbf{Y}(\Theta_1)), \ldots, (\Theta_p, \mathbf{Y}(\Theta_p))$ to obtain predictive distribution for any Θ^*, say $\mathbf{Y}(\Theta^*)$.
2. Add error/discrepancy term to this predictive distribution. This now provides a probability model for the observed summary statistics $\mathbf{Z}$.
3. Inference for $\boldsymbol{\Theta}$: with $\mathbf{Z}$ and above probability model, obtain a likelihood. Can now perform Bayesian (or ML) inference for $\boldsymbol{\Theta}$.

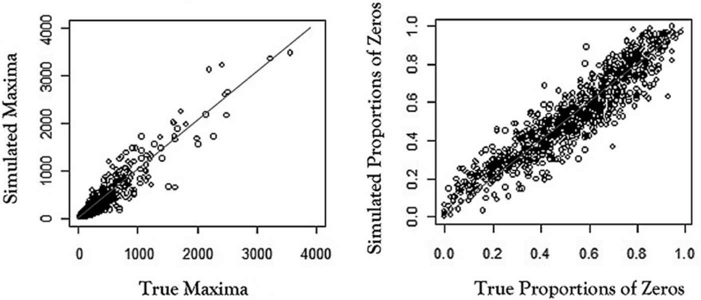
Details: Dimension Reduction

- Space-time data dimensions: 952×546
- Dimensionality of the summary statistics: 952
- # model simulations (# parameter settings): 16,000
- Naive Gaussian process emulation is infeasible
- Solution: emulate distances between summary statistics
  - Using $\mathbf{Y}(\Theta_1), \ldots, \mathbf{Y}(\Theta_p)$ (simulated summary statistics), calculate $d(\Theta_1), \ldots, d(\Theta_p)$, where $d(\Theta_i) =$ distance between $\mathbf{Y}(\Theta_i)$ and $\mathbf{Z}$ (observed summary statistics).
  - Fit a Gaussian process to $(\Theta_1, d(\Theta_1)), \ldots, (\Theta_p, d(\Theta_p))$

Other Details

- Model discrepancy
  - We model the data-model discrepancy as an exponential random variable
  - Accounting for model discrepancy is crucial (Bayarri et al. (2007), Bhat et al. (2010))
- Computational details
  - Slice sampling for fast mixing MCMC algorithm.
  - Gravity model simulations on a $20 \times 20 \times 20 \times 20$ grid are done in parallel on a UNIX cluster.

Model Fit with our Approach

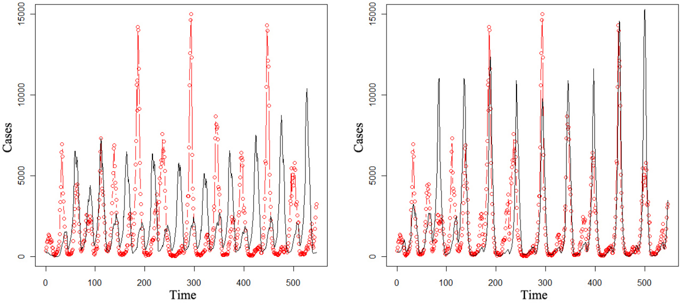


(a)

(b)

Fitted model better captures important biological characteristics

Improved Prediction



(a)

(b)

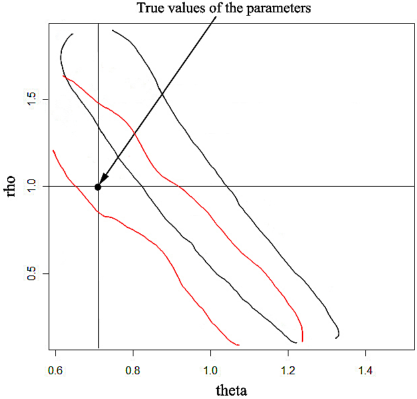
Red: observations, black: predictions

(a) likelihood-based

(b) emulation-based (using summary statistics)

Improved Inference for Θ

95% C.I.'s for (θ, ρ)



Black: likelihood-based

Red: emulation-based (using summary statistics)

Remarks

- Emulation-based inference results in:
  - An improved fit to key biological characteristics
  - Better parameter inference
  - Fast computations
- Unlike previous ad-hoc approaches we can study statistical properties of the gravity TSIR model, including characterizing parameter uncertainty, learning about parameter identifiability issues.
- Biological insights: There are no statistically significant seasonal changes in the movement of the infection between cities.
  - Seasonally forced increase of outbreaks are due to the increase in the local transmission (e.g. in schools)

Summary

- We develop a new Gaussian process-based inferential approach.
  - Focus on summary statistics relevant to the biological phenomena.
  - Scientists build models in order to capture certain key phenomena; makes sense for statisticians to use this information when performing inference for these models.
- Applicable to problems where:
  - the traditional likelihood-based inference is computationally intractable or produces a poor model fit.
  - cost of the simulations from the model make approximate Bayesian computation (ABC) methods infeasible.

Key References

- Xia, Y. C., Bjørnstad, O. N. and Grenfell, B. T. (2004), Measles Metapopulation Dynamics: A gravity model for epidemiological coupling and dynamics, *American Naturalist*.

- Grenfell, B.T., Bjørnstad, O. N. and Kappey, J. (2001), “Traveling waves and spatial hierarchies in measles epidemics.” *Nature*.

- Bhat, K.S., Haran, M., Tonkonojenkov, R., and Keller, K. (2012), “Inferring likelihoods and climate system characteristics from climate models and multiple tracers,” *under revision*

- Bhat, K.S., Haran, M. and Goes, M. (2010) “Computer model calibration with multivariate spatial output,” *Frontiers of Statistical Decision Making and Bayesian Analysis, New York: Springer-Verlag, 2010.*

Emulation-Based Inference for Random Graph Models

- A mixture model for random graphs:
  - Explicitly describes the way edges connect vertices
  - Vertices of the graph spread into Q classes with prior probabilities $(\alpha_1, \cdots, \alpha_q)$
  - Allows for different connectivity probabilities between and within classes
- Networks for metabolic reaction and affiliation networks.
- Current inferential approach: variational methods. Unreliable inference, uncertainty quantification is not straightforward.
- We develop an emulation-based inferential approach (ongoing research).