Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease

Murali Haran[1]

(joint work with Roman Jandarov[1] and Ottar Bjornstad[2])

[1]Department of Statistics, The Pennsylvania State University

[2]Department of Entomology and Center for Infectious Disease Dynamics, The Pennsylvania State University

Joint Statistical Meetings, Vancouver, BC. August 2010

Infectious disease modeling

- Infectious diseases have an immense impact on human health, agriculture and conservation.
- The theory of disease dynamics provides a tractable system for investigating key questions in population and evolutionary biology. Of practical use in management and control of infectious diseases, including immunization and epidemic control strategies.
- SIR models: a class of disease dynamics models.
- Although relatively simple, variants of SIR models have been effective and are widely used.
- Here: statistical inference for the Gravity-TSIR model, which models spatiotemporal dynamics. This model presents several inferential and computational challenges.

SIR models

Basic SIR models classify individuals as one of **susceptible** (S), **infected** (I) or **recovered** (R).

- Individuals are born into the susceptible class.
- Susceptible individuals have never come into contact with the disease and are able to catch the disease, after which they move into the infected class.
- Infected individuals spread the disease to susceptibles, and remain in the infected class (the infected period) before moving into the recovered class.
- Individuals in the recovered class are assumed to be immune for life.

Gravity T-SIR model

- Extension of the discrete time-series SIR (T-SIR) model (Bjornstad et al.2002; Grenfell et al. 2002) with explicit formulation of the spatial transmission between different host communities.
- Notation:
  - $I_{k,t}$ - number of infected individuals in city k at time t.
  - $S_{k,t}$ - number of susceptible individuals in city k at time t.
  - $d_{k,j}$ - distance between cities k and j.
  - $N_{k,t}$ - population of city k at time t.
  - $B_{k,t}$ - local number of new hosts (births) in city k at time t.
  - $L_{k,t}$ - number of infected people moved to city k at time t.
  - T cities, K time points.

Modeling incidences

Following Xia, Bjornstad and Grenfell (2004):

- Number of incidences of a disease at time $t+1$ for city k,

  $I_{k,t+1} = \text{Poisson}(\lambda_{k,t+1})$, where $\lambda_{k,t+1} = \beta_t S_{k,t}(I_{k,t} + L_{k,t})^{\alpha}$.

- $\{\beta_t\}$ specified only via 26 parameters (26 = number of biweeks in a year), to allow for differences in seasonal transmission. Assumed to be same every year.
- $\alpha, \{\beta_t\}$ are local transition parameters.

Modeling susceptibles

- Number of susceptible individuals at time $t+1$ for city k is then modeled via balance equation (Bartlett, 1957):

$$S_{k,t+1} = S_{k,t} + B_{k,t} - I_{k,t+1}$$

- Finally, unobserved number of infected immigrants moved to city k at time t is modeled as:

$$L_{k,t} = \text{Gamma}(m_{k,t}, 1),$$

where

$$m_{k,t} = \theta N_{k,t}^{\tau_1} \sum_{j=1, j\neq k}^{K} \frac{(I_{jt})^{\tau_2}}{d_{k,j}^{\rho}}, \quad \theta, \tau_1, \tau_2, \rho > 0.$$

Statistical inference for measles

- **Measles data**
  - The UK Registrar General's data for 952 cities in England and Wales for years 1944-1966 of biweekly incidences of measles. Very rich spatio-temporal data.
  - Data for number of susceptibles from standard susceptible reconstruction algorithms (Fine and Clarkson 1982a, Schenzle 1984, Ellner et al. 1998, Bobashev et al. 2000, Finkenstadt and Grenfell 2000).
- **Parameters of the model**:
  - Reliable estimates of local transition parameters α and $\{\beta_t\}$ are assumed known from previous work (Bjornstad et al. 2001).
  - Gravity parameters $\theta, \tau_1, \tau_2, \rho$ are unknown.
- **Goal**: Infer unknown gravity parameters: $\theta, \tau_1, \tau_2, \rho$.

Challenges with likelihood-based inference

- Dimensions of the data (TK): 546*952 = 519,792.
- Number of infected immigrants $\{L_{k,t}\}$ are unobserved.
- The likelihood function is complicated:
  - Involves integrating over 519,792 latent variables (frequentist or Bayesian.)
  - Very expensive calculations per iteration.
- Approximate Bayesian computations (ABC) ("MCMC without likelihoods", Marjoram et al., 2002) is also infeasible since simulating draws from this model is computationally expensive.

A simplified model and gridded MCMC

An approach: simplify the model by fixing the number of immigrants (latent variables) at their means.

- Likelihood evaluations are still very expensive. It takes $\approx$ 72 hours to find MLEs for the simplified gravity model.
- Studying likelihood surface, learning about variability of estimates is computationally infeasible.

Gridded Metropolis-Hastings:

- We evaluate expensive parts of the likelihood for a grid of parameter values and store these in a look-up table.
- Sample discretized parameter space (on grid). M-H ratio evaluation is now much faster.

Results

- Simulation experiments and likelihood/Bayes evaluations suggest that the gridded MCMC algorithm produces posterior distributions similar to a non-gridded MCMC algorithm.
- Issues:
  - Serious identifiability issues. Can only infer 2 of the 4 parameters.
  - In simulation studies: posterior (and likelihood) surface is peaked away from the true parameter values. There's a significant bias.

Alternative approach

- Instead of likelihood-based approach, focus on important biological ‘signatures’ (characteristics) of the process. E.g. proportion of zeros (no disease incidences).
- Borrow ideas from computer model emulation, calibration.
  - Simulate realizations from the gravity model at different parameter values.
  - Use the signatures to define statistics.
  - Find distance between summary statistics for the simulated process and the observations.
  - Fit a Gaussian process to this distance.
  - Can obtain a likelihood and perform Bayesian inference for the gravity model parameters using the real data.

Inferential approach outline

- Gravity parameters, $\Theta = (\theta, \tau_1, \tau_2, \rho)$.
- Summary statistics (distance to observations) based on simulations at $\Theta_i, i = 1, \ldots, n$ parameter settings, $\mathbf{Y} = (\mathbf{Y}(\Theta_1), \ldots, \mathbf{Y}(\Theta_n))$.
- Model stochastic model output $\mathbf{Y}$ using a Gaussian process: $\mathbf{Y} \mid \beta, \xi \sim N(\mu_\beta(\Theta), \Sigma(\xi, \Theta))$. Infer β, ξ: regression, covariance parameters.
- Model summary statistic for real data set $\mathbf{Z}$:
- $\mathbf{Z} = \eta(\mathbf{Y}, \theta) + \delta_\Psi(\mathbf{Y}, \Theta) + \epsilon_{\sigma^2}(\mathbf{Y})$
  where η is a random variable with predictive distribution derived above. δ is a discrepancy function, modeled as Gaussian process, and ϵ is a vector of i.i.d. errors.
- Infer posterior $\pi(\Theta, \Psi, \sigma^2 \mid \mathbf{Z}, \mathbf{Y})$ using MCMC.

Summary

- With estimated parameters, the model is able to reproduce well the signatures of the disease process.
- Through simulation studies, we find that our GP-based emulation approach produces unbiased estimates of the parameters.
- Can also produce estimates of uncertainty, look at all marginals, joint distributions of parameters easily.
- Our statistical approach unearths serious identifiability issues.
- Issue: Computational concerns do not allow for more than 3 replications at each grid point.

Key references

- Grenfell, B.T., Bjørnstad, O. N. and Kappey, J. (2001), "Traveling waves and spatial hierarchies in measles epidemics." *Nature*.
- Xia, Y. C., Bjørnstad, O. N. and Grenfell, B. T. (2004), Measles Metapopulation Dynamics: A gravity model for epidemiological coupling and dynamics, *American Naturalist*.
- Bjørnstad, O. N., Finkenstädt, B. and Grenfell, B. T.(2001), Dynamics of measles epidemics. I. estimating scaling of transmission rates using a time series SIR model, *Ecological Monographs*.

Acknowledgment: Support from Bill and Melinda Gates Foundation.