Characterization of Communicable Disease Epidemics using Bayesian Inference

Cosmin Safta\textsuperscript{1}, Jaideep Ray\textsuperscript{1}, Karen Cheng\textsuperscript{2} and David Crary\textsuperscript{2}

\textsuperscript{1}Sandia National Laboratories, Livermore CA, \\
\textsuperscript{2}Applied Research Associates, Inc, Arlington, VA

Presented at the 9\textsuperscript{th} Annual Meeting of the International Society for Disease Surveillance \\
Park City, UT, December 1 & 2, 2010

Ms. Karen Cheng at ARA, is the Principal Investigator. 
Contact Info: kcheng@ara.com and csafta@sandia.gov

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy’s National Nuclear Security Administration under contract DE-AC04-94AL85000.
Outline

- Motivation
- Communicable disease example
- Problem Formulation - Numerical Approach
- Results
  - Parameter Estimation
  - Epidemic Progression
  - Speed-up of the Parameter Inference – Surrogate Models
- Conclusions and Acknowledgement
Motivation & Approach

• Disease models are used in planning, resource allocation etc.
  – They contain parameters which have to be supplied
  – Generally biosurveillance data is used to detect, not characterize outbreaks (some exceptions – Held et al, *Stats. Modelling*, 2005)

• To develop statistical techniques that can characterize an epidemic from biosurveillance data
  – Characterization of the epidemic– estimate number of index cases, (time-dependent) spread rate, etc
    • NOT trying to characterize the pathogen – no genetic, immune-system response, etc.
  – Use biosurveillance data and real-time estimation
    • Estimates will be highly uncertain, so need to quantify uncertainty

• Questions
  – How small an epidemic can we detect and characterize?
  – What can we characterize with useful uncertainty bounds?
A Communicable Disease Example

• A simulated plague epidemic
  – Performed with an agent-based model for disease spread; includes visit-delay
    • Disease parameters from Gani & Leach, *EID*, 2004
    • Insert into ICD-9 stream for ILI from Miami
    • 100/1000 index cases; epidemic dies out in 40 days

• Extract epidemic, per Ray et al, *CBD Conf*, Orlando, 2010

• Aim:
  – Estimate the total size of the epidemic
  – Also, the infection rate and visit delay curves
  – Compare with the “true” figures from the simulation
Extraction of the Epidemic

100 Index Cases

1000 Index Cases

- No. of days since start of epidemic
- No. of people reporting

- Reported data
- KF prediction
Formulation of the Problem

• **Data** – the extracted epidemic: time-series of counts of people seeking care, on a daily basis

• **Model** - A *convolution* of a time-dependent infection rate (1 free parameter), incubation period (known), and *visit delay* (1 free parameter)
  
  – Also includes total size of the epidemic, time of infection of the index cases and fraction of index cases as free parameters (Brookmeyer’s 1988)

  – 5 free parameters in all

• **Fitting**
  
  – Estimate the PDFs of the 5 parameters using an adaptive Markov Chain Monte Carlo (MCMC) approach

  – Takes about 1-3 hrs depending upon the length of the time-series
Estimation of the No. of Index Cases

- The true values are 100 and 1000, respectively

- The estimate improves with time (and data!) for larger outbreaks

- Estimates performed with data starting from
  - Start of epidemic + 4 (s+4)
  - Start of epidemic + 6 (s+6)

- Easier for large outbreaks
Estimation of the Start of the Epidemic and its Total Size

Epidemic starts 4 and 6 days, respectively, before data collection.

Total size true value is approx. 11000.

[Graphs showing data analysis]
Estimation of the Parameters in Infection Rate and Visit Delay Models

- Both modeled as a $\Gamma$-functions
  - rate parameters are inferred; shape parameters are set
Joint Probability Distributions of the Inferred Parameters

5 Days of Data

15 Days of Data
Estimation of the Epidemic’s Progression

- Best estimate – based on maximum a-posteriori (MAP) distribution
- Developed using 15 days of data, starting 4 and 6 days, respectively, after first 1000 people got infected
Speed up the Inference – Surrogate Models

\[ v_{\text{ind}} \left( (t_i, t_{i+1}] \right) = N_{\text{tot}} (1 - \alpha) \int_{\tau}^{t_{i+1}} f_{\text{inc}} (s - \tau) \left[ F_{\text{vd}} (t_{i+1} - s; r_{vd}) - F_{\text{vd}} (t_i - s; r_{vd}) \right] ds \]

\[ v_{\text{sec}} \left( (t_i, t_{i+1}] \right) = N_{\text{tot}} \alpha \int_{w=\tau}^{t_{i+1}} \int_{u=\tau}^{t_{i+1}} q_{\text{inf}} (u - \tau; r_{ir}) f_{\text{inc}} (w - u) \left[ F_{\text{vd}} (t_{i+1} - w; r_{vd}) - F_{\text{vd}} (t_i - w; r_{vd}) \right] du dw \]

\[ v_{\text{tot}} = v_{\text{ind}} + v_{\text{sec}} \]

- Double integral is very costly during the MCMC sampling
- Create a surrogate for the epidemic model and compute it offline
  - Use Polynomial Chaos representations (Ghanem & Spanos, 1991): accurate with respect to the pdf’s of interest and fast to evaluate.

\[ v_{\text{tot}} (t) = \sum_{k=1}^{P} a_k \Psi_k^{(6)} (N_{\text{tot}}, \alpha, \tau, r_{vd}, r_{ir}, t) \rightarrow v_{\text{tot}} (t - \tau) = N_{\text{tot}} \sum_{k=1}^{P} (b_k + c_k \alpha) \Psi_k^{(3)} (r_{vd}, r_{ir}, t) \]
Surrogate Models – cont’d

Visit delay rate = 0.2
• Red mesh ➜ direct model
• Blue mesh ➜ surrogate model

Visit delay rate = 1.0
Conclusions

• Early in the development of techniques to characterize epidemics
  – Working off biosurveillance data
  – Provides information on the particular/ongoing outbreak
  – Second half of a detect-and-characterize algorithm; model selection algorithm is also in place

• Parameter estimation capability ideal for providing the input parameters into an agent-based model
  – Index cases, spread/infection rate, total epidemic size, etc
  – Since it’s real-time, can be used to check if medical interventions are effective

• To do
  – Tests with different kinds of background models
  – Tests with outbreaks of different sizes and spread/infection rates
  – Identification of a “proper” set of ICD-9 codes for monitoring biosurveillance data streams
Acknowledgements

The work was funded by DTRA under contract HDTRA1-09-C-0034

Dr. Nancy Nurthen is the DTRA PM.

Ms. K. Cheng at Applied Research Associates, Inc, is the PI