Statistical Science
2018, Vol. 33, No. 1, 1–3
https://doi.org/10.1214/18-STS331I
© Institute of Mathematical Statistics, 2018

Introduction to the Special Section on Inference for Infectious Disease Dynamics

Theodore Kypraios and Vladimir N. Minin

1. INTRODUCTION

The past three decades have seen significant growth in the field of mathematical modeling of infectious diseases, leading to substantial increase in our understanding of epidemiology and control of these diseases. Ability to quickly unravel the dynamics of the spread of infectious diseases is important for effective prevention of future outbreaks and for control of ongoing ones. Recent interest in infectious disease modeling was initially stimulated by the discovery of HIV in the early 1980s and has been maintained by the need to respond to other infectious disease related crises such as, for example, foot-and-mouth disease and SARS outbreaks, healthcare-associated infections, and elevated risks of human influenza pandemics (e.g., risks of global spread of avian and swine flu). Recent Ebola and Zika outbreaks further underscored importance of mathematical and statistical analyses of epidemic dynamics. As a result, mathematical infectious disease modeling remains high on the global scientific agenda.

To respond to challenges posed by infectious diseases the epidemic modeling community has become very engaged in public health policy development, further stimulating interest in models of disease transmission. This involvement in practical applications of infectious disease modeling was enabled by algorithmic advances and the continuing increase in computing power. For example, it is now possible to perform simulations based on parameter rich and realistic *agent-based models* that generate individual-level behavior, infections, and recoveries of millions of individuals. Moreover, analyses on infectious disease outbreak data can now be performed using computationally intensive methods such as maximum likelihood estimation, Markov chain Monte Carlo (MCMC), sequential Monte Carlo (SMC), and approximate Bayesian computation (ABC).

The transmissible nature of infectious diseases makes them fundamentally different from noninfectious diseases, and therefore it is difficult to analyze disease outbreak data using off-the-shelf statistical methods. This is mainly due to (i) *presence of strong and complex dependencies in the data* and (ii) *missing data significantly surpassing observed data in size*, because the actual transmission process cannot be directly observed. Therefore, specialized and problem-specific methods are required. Despite recent significant advances in fitting stochastic epidemic models to data, there are still a number of challenges to overcome.

In view of the ever growing research activity in the area and the practical importance of effectively analyzing disease outbreak data, we have organized this special section of *Statistical Science* in which the aim is to provide an overview of the most recent developments as well as the current research challenges facing the epidemic modeling community. We hope that readers of the journal will get a broad idea of the field as well as of its current research directions.

2. STOCHASTIC EPIDEMIC MODELS

There exist many forms of models to represent the dynamics of infectious diseases. Most models are concerned with a population consisting of individuals who are potentially able to transmit the disease to one another. In this section we focus on stochastic models. Although in some settings deterministic models can approximate stochastic dynamics well, behavior of stochastic and deterministic models can be qualitatively different in certain parameter regimes (e.g., when the disease is spreading in a small population). This suggests that stochastic models should be preferred to deterministic ones if one wishes to stay true to the stochastic nature of infectious disease dynamics.

Stochastic epidemic models are generally defined at the level of individuals, for instance, specifying probability distributions that describe how long an individual remains infectious. A key aspect of any disease

Theodore Kypraios is Associate Professor, School of Mathematical Sciences, University of Nottingham, NG7 2RD, United Kingdom (e-mail: theodore.kypraios@nottingham.ac.uk). Vladimir N. Minin is Professor, Department of Statistics, University of California, Irvine, California 92617, USA (e-mail: vminin@uci.edu).

transmission model is the set of assumptions made regarding transmission itself, that is, the mechanism via which susceptible individuals become infected. Let us consider a continuous-time *Susceptible-Infective-Removed* (SIR) model, defined as a continuous-time Markov chain whose state space is the numbers of infected and susceptible individuals. The SIR model is important because all mechanistic models of infectious disease dynamics can be thought of as extensions of the SIR model, yet this simple model fully illustrates statistical challenges arising during analyses of stochastic epidemic models. The SIR model typically assumes homogeneous mixing of individuals in the population, leading to the law of mass action that says that the rate of adding a new infection to the population is a product of the numbers of susceptible and infected individuals and the infection rate β. Furthermore, all individual infectious periods are independent exponentially distributed random variables with mean $1/\gamma$. Equivalently, the rate of introducing a new recovery/removal in the population is γ times the number of currently infected individuals. The SIR model has two parameters of interest, namely the infection rate β and the removal rate γ. If both the times at which individuals got infected and removed where known, drawing inference for β and γ would be trivial. However, in practice, disease symptoms develop well after the time of infection and, as a result, infection times are almost always unknown. Consequently, infectious disease outbreak data are always incomplete, leading to a computationally intractable likelihood of the observed data (e.g., removal times). State-of-the-art methods deal with this intractability using three main strategies: (1) maximum likelihood estimation, where missing data are integrated out with the help of SMC; (2) Bayesian data augmentation with MCMC (possibly with embedded SMC) targeting the joint distribution of model parameters and missing data; (3) ABC approach.

As mentioned before, the SIR model is a basic building block in stochastic epidemic modeling. There are two main ways in which the basic SIR model can be extended. First, the number of compartments and their definitions can be tailored to a particular application. For example, effects of demographic variables (e.g., age and sex) on disease transmission can be modeled by creating S, I, and R compartments of multiple types. The second way to extend the SIR model and its relatives is to make infection rate and possibly other parameters time dependent. Such time inhomogeneous models are used to account for seasonality and to incorporate effects of time varying environmental factors on infectious disease dynamics. We will see both of these extension types in the special section papers.

3. ARTICLES IN THIS SECTION

This special section consists of six articles which present a general overview of the state-of-the-art in a number of different topics concerning stochastic epidemic models for infectious disease data.

- McKinley, Vernon, Andrianakis, McCreesh, Oakley, Nsubuga, Goldstein and White ("Approximate Bayesian Computation and Simulation-Based Inference for Complex Stochastic Epidemic Models") first provide an overview of the more popular variants of ABC and then discuss some of the challenges that one is faced with when applying ABC to high-dimensional and computationally intensive models. They then discuss an alternative approach—history matching—that aims to address some of these challenges and provide a comparison between the two different approaches.
- Gibson, Streftaris and Thong ("Comparison and Assessment of Epidemic Models") consider a variety of stochastic representations of individual-based continuous-time epidemic models and review the range of model-comparison and model-assessment approaches that are currently available. In particular, they highlight some of the factors, such as lack of replication, partial observation of processes and the nonnested nature of models to be compared, that can impede checking and criticism of epidemic models.
- Birrell, De Angelis and Presanis ("Evidence Synthesis for Stochastic Epidemic Models") provide an overview of evidence syntheses in stochastic epidemic modeling where multiple types of data are explicitly used in an integrated analysis. The authors discuss recent developments in this area and highlight the ongoing and future challenges, such as potential of conflicting evidence as well as computationally efficient methods for inference.
- Kypraios and O'Neill ("Bayesian Nonparametrics for Stochastic Epidemic Models") provide an overview of Gaussian process-based Bayesian nonparametrics applied to stochastic epidemic models. In particular, the authors concentrate on estimating changes of the per capita infection rate across time and on replacing the law of mass action with nonparametric estimation of the rate of infection.
- Bretó ("Modeling and Inference for Infectious Disease Dynamics: A Likelihood-based Approach") gives an overview of likelihood-based methods that

allow for estimation of stochastic epidemic model parameters. The author emphasizes maximum likelihood estimation and discusses importance of hierarchical modeling to account for per capita infection and recovery rate heterogeneity within the population of interest.

- Kendall, Ayabina and Colijn ("Estimating Transmission from Genetic and Epidemiological Data: A Metric to Compare Transmission Trees") review methods that estimate transmission ("who infected whom") trees. The authors also present a new metric that quantifies differences in two transimission trees and demonstrate how this metric helps interpreting the posterior distribution of transmission trees in Bayesian analysis.

4. RESEARCH OUTLOOK

In spite of great progress in statistical analyses of infectious diseases, many challenges remain. The majority of the state-of-the-art methods for fitting stochastic epidemic models to data do not scale well with model and/or data complexity. For example, working with epidemics spreading in large populations with many compartments remains challenging. Several articles in the special section address these challenges, but developing efficient algorithms for fitting stochastic epidemic models will remain an active research topic in the foreseeable future.

There has been significantly more focus on developing methods for fitting stochastic epidemic models than on methods for model assessment. Although there exist methods that in principle enable discrimination between competing models, using, for example, Bayes factors, their implementations are often problem-specific and their adoption by the practitioners is slow. One future challenge is to promote use of formal model comparison among epidemiologists. Another important goal is to develop methods for assessing the goodness of model fit—a topic discussed by one of the papers in the special section.

One of the papers in the special section discusses integration of multiple data sources. An emerging example of such integration is joint use of epidemiological and genetic data. As genetic sequencing technologies become affordable and accessible, genetic data get used routinely in infectious disease surveillance and during responses to sudden outbreaks. However, statistical methods capable of integrating epidemiological and genetic data in a fully probabilistic framework still face significant computational challenges. Although our special section papers do not review methodological state-of-the-art of inferring infectious disease dynamics from genetic data, one of the papers provides a useful overview of available methods for inferring transmission networks from genetic data and discusses challenges in epidemiologically meaningful interpretation of results produced by these methods. We predict that use of genetic data and, more generally, data integration will remain an important research theme in infectious disease epidemiology.