**SANDIA REPORT**

SAND2011-6776
Unlimited Release
Printed September 2011

# Real-time Characterization of Partially Observed Epidemics using Surrogate Models

Cosmin Safta, Jaideep Ray, Khachik Sargsyan, Sophia Lefantzi, Karen Cheng, and David Crary

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.



Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.



Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2011-6776
Unlimited Release
Printed September 2011

# Real-time Characterization of Partially Observed Epidemics using Surrogate Models

Cosmin Safta, Jaideep Ray, Khachik Sargsyan, Sophia Lefantzi
Sandia National Laboratories, Livermore, CA
{csafta,jairay,ksargsy,slefant}@sandia.gov
and
Karen Cheng, David Crary
Applied Research Associates, Arlington, VA
{kcheng,dcrary}@ara.com

**Abstract**

We present a statistical method, predicated on the use of surrogate models, for the "real-time" characterization of partially observed epidemics. Observations consist of counts of symptomatic patients, diagnosed with the disease, that may be available in the early epoch of an ongoing outbreak. Characterization, in this context, refers to estimation of epidemiological parameters that can be used to provide short-term forecasts of the ongoing epidemic, as well as to provide gross information on the dynamics of the etiologic agent in the affected population e.g., the time-dependent infection rate. The characterization problem is formulated as a Bayesian inverse problem, and epidemiological parameters are estimated as distributions using a Markov chain Monte Carlo (MCMC) method, thus quantifying the uncertainty in the estimates. In some cases, the inverse problem can be computationally expensive, primarily due to the epidemic simulator used inside the inversion algorithm.

We present a method, based on replacing the epidemiological model with computationally inexpensive surrogates, that can reduce the computational time to minutes, without a significant loss of accuracy. The surrogates are created by projecting the output of an epidemiological model on a set of polynomial chaos bases; thereafter, computations involving the surrogate model reduce to

evaluations of a polynomial. We find that the epidemic characterizations obtained with the surrogate models is very close to that obtained with the original model. We also find that the number of projections required to construct a surrogate model is $O(10) - O(10^2)$ less than the number of samples required by the MCMC to construct a stationary posterior distribution; thus, depending upon the epidemiological models in question, it may be possible to omit the offline creation and caching of surrogate models, prior to their use in an inverse problem. The technique is demonstrated on synthetic data as well as observations from the 1918 influenza pandemic collected at Camp Custer, Michigan.

# Acknowledgment

This work was supported by the DTRA Contract HDTRA1-09-C-0034. We want to acknowledge helpful suggestions from Habib Najm in the development of this report. Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94-AL85000.

This page intentionally left blank.

# Contents

# List of Figures

# List of Tables

This page intentionally left blank.

# Chapter 1

# Introduction

Epidemiological models, i.e., models that predict the evolution of an epidemic, given certain model parameters, are routinely used to characterize diseases from outbreak data. Often, these are used in retrospective studies to estimate epidemiological variables which form the model inputs. The rate of spread of a communicable disease is a commonly estimated model parameter [13, 5, 7, 16, 43]; the genesis of outbreaks caused by accidents [6, 58, 40] is another example. Fewer studies have targeted the use of models to estimate in real-time the probability of human transmission for emerging infectious diseases [3, 60] or to gauge the effect of countermeasures in an ongoing outbreak [51, 34, 54].

The estimation of epidemiological parameters, in real-time, pose certain challenges[1]. The data is generally sparse and often, only hospitalization times are available, rather than times of appearance of symptoms. Thus models, fitted to real-time data, have to account for the hospital visit delay [51]. Further, all estimates are generally uncertain and estimates are best expressed as distributions [34] developed via a Monte Carlo sampler. A particular difficulty faced during online model fitting to data, especially during the early stages of the outbreak, is the representation of the highly variable dynamics associated with disease spread; while sophisticated modeling may be able to address these, the computational expense of Monte Carlo sampling does not allow their use within time-constraints of online estimation. Thus most epidemiological models are compartmental ones using some variant of uniform mixing to model spread, though modified by a time-dependent effective reproduction number [43, 51]. Clearly, an ability to reduce the computational time of a disease model can favorably impact the fidelity with which an outbreak can be characterized from partial observations.

In this paper, we demonstrate a method to do so. At its core, it reduces to replacing the epidemiological model with a polynomial surrogate, which can be made arbitrarily accurate (at the expense of computational cost). The choice of the variable being modeled by the surrogate is crucial; smoothly varying functions are easily approximated by parsimonious surrogates. The surrogate model is created by projecting the output of the epidemiological model, run repeatedly with a sampled set of input parameters, on a basis set; a weighted sum of the bases constitutes the surrogate model. The bases are chosen to minimize the number of model evaluations and maximize the fidelity with which the resultant surrogate reproduces the original model. However, the replacement

[1]Note that in epidemiology, where data is often available only on a daily resolution, a "real-time" computational process is defined as one that can accomplished in considerably less than a day - for our purposes, we take it to mean less than an hour.

of the "true" epidemiological model with a surrogate in the parameter estimation problem introduces an error in the inferred parameters and we explore the magnitude and nature of its impact; in principle, the impact of the model error can be made small enough so that it is negligible compared to the errors due to lack of data or due to imperfect measurements. We also investigate the efficiency gained, as measured by the reduction in computational time, by employing the surrogate instead of the original model. The cost of building the surrogate model in the first place is also included in this analysis.

The rest of the paper is organized as follows. In Chapter 2 we present a literature review of existing work on the estimation of partially observed epidemics and the construction of surrogate models using polynomial chaos expansions. In Chapter 3 we formulate an inverse problem for the characterization of epidemics with partial observations, describe the epidemiological models used in the inverse problem and detail the method by which synthetic epidemiological data (used later in tests) was generated. We also solve the inverse problem, and develop estimates of epidemiological parameters using an adaptive Markov chain Monte Carlo (MCMC) method. In Chapter 4, we describe the method to construct the surrogate model and recompute the estimates obtained in Chapter 3 using surrogates. The differences in the epidemiological estimates so obtained (vis-à-vis Chapter 3) are quantified, along with the savings in computational time. We conclude in Chapter 5.

# Chapter 2

# Literature Review

In this section we review existing literature on the estimation of epidemiological parameters as well as the use of polynomial chaos expansions to construct computationally inexpensive surrogate models. The former will focus on methods that are amenable to be used in a real-time setting, where only partial observations may be available.

## 2.1 Estimation of epidemiological parameters from partial observations

Real-time estimation of epidemiological characteristics, using time-series data from an on-going outbreak, has recently gained prominence. Most of the methods have targeted the estimation of a time-dependent spread-rate, often couched in terms of their effective reproductive number $R_t$. In [2] Bettencourt describes a statistical method based on sampling a prior distribution of epidemiological model parameters, and iteratively forming a posterior distribution based on comparing simulated epidemic evolutions to sparse observations, with a view of improving the predictive skill of the model. His earlier paper [3] developed a Bayesian technique to estimate a time-dependent $R_t$ (for various influenza outbreaks), conditioned on streaming data. In [42] Nishiura *et al.* develop an epidemic model that includes a time-dependent $R_t$, and an estimator for it based on the serial interval observed in an outbreak. The model was fit to historical data.

Real-time epidemiological characterization can also be done using data from contact tracing. Wallinga and Teunis [54] developed a method, based on contact tracing data, to estimate the $R_t$ for SARS outbreaks in Hong Kong and elsewhere and gauge the impact of countermeasures on the outbreaks. The method is purely retrospective, requiring full knowledge of chains of transmission, and is similar to the work (done for plague outbreaks) in [43]. Cauchemez *et al.* [8, 9] adapted the method to be applicable in a real-time context, where data on only a small sample of transmission chains and a small number of symptomatic secondary cases are available. The model assumes that no index cases are injected into the population after the start of the epidemic and there is no delay between the appearance of symptoms and hospitalization time. They developed posterior distributions for $R_t$ for the SARS epidemic as data became available; within 25 days of the start of the epidemic (and 5 days post implementation of countermeasures), $R_t$, as estimated from data from Hong Kong, showed an exponential decline. A similar decline was calculated for plague outbreaks

in [22]. More recently, particle filters have been used to provide forecasts of H1N1 outbreaks in 2009 in Singapore [45]. In [51, 34] the authors track the 2002 SARS epidemic in Hong Kong with a compartmental epidemic model where disease transmission due to superspreading and non-superspreading events were represented separately. They developed estimates (distributions) of $R_t$. A novelty in their approach was the inclusion of a model for visit delay, i.e., unlike the work described above, they did not assume that the time of exhibition of symptoms was known; the data consisted of the times that symptomatic patients sought care.

Retrospective methods to estimate spread-rate of a disease, in the face of partial data and structured populations have also been demonstrated for influenza [10] and smallpox [13]. A very different approach was followed in [5, 48] where they inferred the spread rates and the chains of transmission over a latent social network. The approach was Bayesian and distributions for the estimated quantities were developed. Brookmeyer and colleagues [32, 7] have developed a method to estimate a latent time-dependent infection rate by convolving it with the incubation period distribution and equating it to noisy observations. Smoothness constraints were imposed on the time-dependent infection rate profile, and a point estimate (i.e., no uncertainty bounds) was obtained by expectation maximization. The method was used to estimate the evolution of the infection rate of HIV in the 1980s and 1990s in USA and provide forecasts of disease incidence.

Far less work has been done in the estimation of epidemics caused by non-communicable diseases. They mostly deal with anthrax epidemics [50, 31, 53, 58, 25], caused either through an attack or an accidental release. In [50, 31, 53] the authors employ a Bayesian formulation to pose an inverse problem to infer the time of the attack, its location, dosage, the number of index cases and their distribution in space with application to prioritizing the care of the infected people. A time-series less than a week long was sufficient to draw inferences which were informative enough to mount a response. The inference was in the form of a distribution for the estimated quantities. In [25] a slightly different approach was followed, not to characterize an anthrax attack but to provide an alarm (via syndromic surveillance) under the assumption that an anthrax attack had occurred. Nevertheless, the procedure required one to estimate the size of the attack, which followed a Bayesian formulation but obtained the estimates via maximum likelihood estimation. The work in [40] analyzed the Sverdlovsk accident by fitting a model of aerosol dispersion to the residential locations of the approximately 70 people infected in the accident; the fit showed that the location of the release was a military compound where anthrax was used for medical research. This may be considered to be an early (and manual) approach to the characterization techniques described in [31, 25]. In [58], the approach outlined in [40] was followed to elucidate the dose-dependent incubation period of the anthrax.

In this paper, we will extend Brookmeyer's approach [7] so that it can be used in a real-time setting, with data that reflect symptomatic patients seeking care at healthcare facilities. We do so by augmenting it with a model for visit delay. Unlike Brookmeyer, we will assume a parametric form for the infection rate; furthermore, the form will allow for the introduction of index cases into the affected population at arbitrary times. This allows the introduction of transient index cases e.g., travellers, who can seed a transmission chain in a population without contributing to the morbidity time-series obtained from it. In doing so, we partially relax some of the assumptions inherent in Cauchemez *et al.* 's construction in [8, 9].

## 2.2 Surrogate models using polynomial chaos expansions

Surrogate models are computationally inexpensive analogs of expensive computational models. These models approximate one (and sometime more) outputs of a model as a function of model inputs. The surrogate models, sometimes also called response-surface models, typically do not have any scientific/phenomenological arguments underlying their construction and can be likened to "curve-fitting". The primary issues involved during surrogate model construction are (1) minimizing the number of expensive-model evaluations to generate the data to which surrogates are fit (generally accomplished with some kind of sampling) and (2) minimizing the difference between predictions/outputs of the expensive model and its inexpensive surrogate. Descriptions of the issues involved in generating surrogates can be found in [46, 44, 24, 14]. Surrogate models are popular in inversion and optimizations studies since they involve repeated evaluation of models for different parameter values.

In this work we will employ polynomial chaos (PC) expansions to construct surrogate models that will replace the costly epidemic model evaluations during the inference process. The polynomial chaos (PC) was defined first by Wiener [57], and it has since found a significant number of applications in various engineering fields [19, 17, 18, 59, 12]. This approach consists of approximating a generic random variable in terms of standard random variables through a spectral polynomial expansion. In the context of this paper the disease evolution will be cast as a random variable that is function of uncertain input parameters that define the epidemiological models. These spectral approximations are constructed using a relatively small number of function evaluations, and can represent accurately the smooth input-output dependencies. For cases where the model exhibits non-smooth behavior, several domain partitioning methods have been proposed [28, 55, 52]. This generates a series of sub-domains where models have a smooth behavior, thus enabling the use of efficient spectral approximations in each region.

Marzouk *et al.* [38] proposed using surrogate models based on PC expansions in order to accelerate Bayesian inferences. This approach was followed by several authors in a wide range of scientific fields; for source and parameter estimation in porous media [35, 33], analysis of supersonic combustion [11], stochastic data assimilation [39] to name a few. To our knowledge this work is the first attempt to accelerate the inference of epidemic model parameters using a surrogate model approach based on PC representations.

This page intentionally left blank.

# Chapter 3

# Statistical Characterization of Partially Observed Epidemics

In this section, we formulate a Bayesian inverse problem to estimate epidemiological parameters conditioned on sparse data. The data consist of a truncated time-series of symptomatic patients diagnosed with the disease, collated on a daily basis, as might be available in the early epoch of an epidemic. The time a patient seeks care at medical facilities is used for data collation (rather than time of appearance of symptoms) since this information is generally easily available. We also discuss the epidemic models used in the inverse problem and the sources of data (both real and synthetically generated). We conclude with a demonstration of the approach on three different outbreaks and investigate the length of the time-series of observations required to estimate the epidemiological parameters to a given level of accuracy.

## 3.1 Formulation of Inverse Problem

Consider an epidemic seeded by $N_{index}$ index cases. The stream of symptomatics, $\nu_{tot}$, reporting for care in an interval $[t_i, t_{i+1})$ consists of two parts (a) $\nu_{ind}$ , number of symptomatic people that were index cases, and (b) $\nu_{sec}$, number of symptomatic people that were not index cases, i.e. they were infected subsequently as the disease spread. Here, $\Delta t = t_{i+1} - t_i$ is usually 1 day.

The index-case component, $\nu_{ind}$, observed in $(t, t+\Delta t)$ can be given by

$$\nu_{ind} = N_{tot}\,(1-\alpha)\int_{\tau}^{t_{i+1}} f_{inc}(s-\tau;\theta_{inc})\,[F_{vd}(t_{i+1}-s;\theta_{vd}) - F_{vd}(t_i-s;\theta_{vd})]\,ds \qquad (3.1)$$

where $N_{tot}$ is the total number of people infected during the course of the epidemic (i.e., the final size of the epidemic), and $\alpha$ is the fraction of people showing symptoms that are not index cases. For the index cases the incubation starts at the time of the infection $\tau$. The probability of developing symptoms between time $s$ and $s+ds$ is given by $f_{inc}(s-\tau;\theta_{inc})$, (where $f_{inc}$ is the probability density function for the incubation period) and $F_{vd}(t_i-s;\theta_{vd})$ is the cumulative distribution function (CDF) for the visit delay. Here $\theta_{inc}$ and $\theta_{vd}$ are parameters that control the incubation period and visit delay models, respectively. Note that we have used the fact that $F_{vd} = 0$ for $(t-s) < 0$ to simplify the above expression. The models for the incubation period and the visit delay are in Sec. 3.2.

The number of secondary cases, $\nu_{sec}$, is given by

$$\nu_{sec} = N_{tot}\alpha \int_{w=\tau}^{t_{i+1}} \int_{u=\tau}^{t_{i+1}} q_{ir}(u-\tau;\theta_{ir})\, f_{inc}(s-\tau;\theta_{inc}) \times \\ [F_{vd}(t_{i+1}-w;\theta_{vd}) - F_{vd}(t_i-w;\theta_{vd})]\, du dw \tag{3.2}$$

For the secondary cases, the infection takes place at time $u$ according to the infection rate modeled by $q_{ir}$. The visit delay is also applied to the secondary cases and, together with the infection model, results in the double integral above. The infection rate model, which depends on parameter $\theta_{ir}$, is described in Sec. 3.2.

Thus, the total number of people requesting medical care in the interval $(t_i, t_{i+1})$ is given by the sum

$$\nu_{tot} = \nu_{ind} + \nu_{sec} \tag{3.3}$$

and depends on the set of parameters $\Theta = (N_{tot}, \alpha, \tau, \theta_{inc}, \theta_{vd}, \theta_{ir})$. Here the incubation period, visit delay, and infection rate models, can be controlled by one or more parameters.

Given data $\mathbf{d}$ in the form of a time-series of observed $\nu_{tot}(t_i, t_{i+1}]$, the epidemic model parameters $\Theta$ can be estimated in the form of a multivariate PDF via Bayes theorem:

$$p(\Theta|\mathbf{d}) = \frac{p(\mathbf{d}|\Theta)\cdot p(\Theta)}{p(\mathbf{d})} \tag{3.4}$$

where $p(\mathbf{d}|\Theta)$ is the probability distribution of observing the data $\mathbf{d}$ (also called the likelihood function), given a particular $\Theta$, $p(\Theta)$ is our prior belief distribution in that particular value of $\Theta$, $p(\mathbf{d})$ is the probability of observing the data. This term is a normalization factor in Eq. 3.4 and is not important when computing $p(\Theta|\mathbf{d})$ which is the posterior distribution of $\Theta$ conditioned on $\mathbf{d}$. The likelihood $p(\mathbf{d}|\Theta)$ describes the discrepancy, here assumed Gaussian, between the number of symptomatic people predicted by the model and the number of symptomatic people observed:

$$p(\mathbf{d}|\Theta) = \prod_{i=1}^{N_d} \exp\left(-\frac{\left(\nu\left((t_i, t_{i+1}]\right) - n_i\right)^2}{2\sigma_M^2}\right) \tag{3.5}$$

where $\{n_i, i = 1, \ldots, N_d\}$ is the time series of symptomatic people requesting medical care. The standard deviation, $\sigma_M$, between the model and observations can also be inferred along with the model parameters. However its value does not affect the conclusions of this paper on the use of surrogate approximations to replace the expensive epidemiological models. For this reason, we chose a constant value, $\sigma_M = 150$, for all results presented in this paper.

A Markov Chain Monte Carlo (MCMC) algorithm is used to sample from the posterior probability $p(\Theta|\mathbf{d})$. MCMC is a class of techniques that allows sampling from a posterior distribution by constructing a Markov Chain that has the posterior as its stationary distribution [15, 20]. In particular, we use an adaptive Metropolis algorithm [23]. This methodology is an improvement over the original Metropolis algorithm [41]. It uses the covariance of the previously visited chain states to find better proposal distributions, allowing it to explore the posterior distribution in a far more efficient manner; see [23] for details.

## 3.2 Epidemic Models and Priors

We describe the models, specifically the PDFs and CDFs used to describe epidemiological variables. We will do so for the etiologic agents (plague, influenza and anthrax) to be used in this study.

The incubation period is described using a log-normal distribution i.e.

$$f_{inc}(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{\log^2(t/\mu_D)}{2\sigma^2}\right) \tag{3.6}$$

For plague the values for $\mu_D$ and $\sigma$, obtained from [16], are 4.3 and 0.3762, respectively. For influenza, the corresponding figures are $\{1.79, 0.47\}$, indicating a mean incubation period of 2 days and a variance of 1 day [4]. For anthrax, $\mu_D$ is dose dependent and is obtained from Wilkening's A2 model [58]. $\sigma = 0.804 - 0.079 \cdot \log_{10} D$, is also taken from [58], where $D$ is the dose, in terms of spores inhaled by the infected individual. For the anthrax model, $\theta_{inc}$ is set to $\log_{10} D$.

The visit delay i.e., the delay between exhibition of symptoms and the time at which a symptomatic seeks care, is modeled using a Gamma function. A log-normal model for the visit delay observed for severe diseases is available in [26], based on data collected by [27]. However, since the log-normal model was used in the epidemic simulators employed to generate synthetic data for our tests, we adopted a Gamma model in the inverse problem to prevent an "inverse crime" (using the same model to generate the synthetic data and then infer the parameters). The CDF for the visit delay is given by

$$F_{vd}(t;\theta_{vd}) = \frac{\theta_{vd}^{1.992}}{\Gamma(1.992)} \int_0^t \tilde{t}^{0.992} \exp(-\theta_{vd} \cdot \tilde{t})\, d\tilde{t}, \tag{3.7}$$

In this equation, the shape parameter, 1.992, is obtained by fitting to the log-normal model in [26]. However, the rate parameter, $\theta_{vd}$ is left as an unknown (i.e., to be inferred when solving the inverse problem) since the visit delay can shorten during an outbreak as the population becomes aware of it.

We model the rate at which the secondary cases are infected using a Gamma distribution. This is best conceived as the number of people infected on a daily basis since the time of introduction of the index cases. The Gamma function, for appropriate parametric values can model an epidemic when the infection rate initially increases (as more infectious people become available in the population) followed by a waning as countermeasures are put in place. The peak of the infection rate and the speed of its decay can be adjusted parametrically. The initial rate of increase is controlled by its shape parameter $k$ which is generally difficult to infer from partial data. We use $k = 2$ for plague [49] and $k = 23$ for influenza (see [30] for derivation).

$$q_{ir}(t;\theta_{ir}) = \frac{\theta_{ir}^k}{\Gamma(k)} \int_0^t \tilde{t}^{k-1} \exp(-\theta_{ir} \cdot \tilde{t})\, d\tilde{t}, \tag{3.8}$$

Here $\theta_{ir}$ is a rate parameter that largely controls the decay of the spread (infection) rate. Since this decay will be affected by medical countermeasures, we leave this as a parameter to be inferred from data.

The priors used in the inference of epidemiological model parameters will be defined in Sec. 3.4.

## 3.3 Epidemic Data

The time-series data **d** used in this study is generated using epidemic simulators (for plague and anthrax outbreaks) and actual observations from the 1918 influenza pandemic. The methodology for simulating epidemics caused by an aerosol release is described in detail elsewhere. We reproduce a summary below.

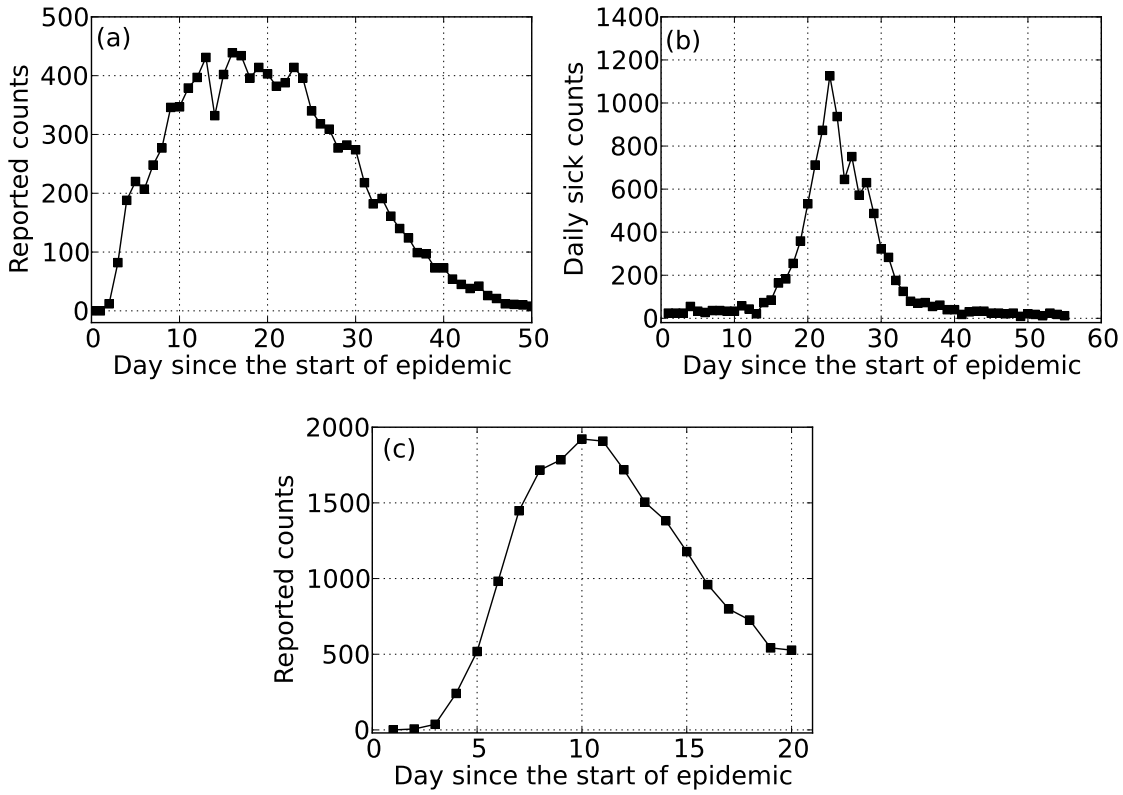


**Figure 3.1.** Time series of reported counts of symptomatic cases for (a) plague, (b) 1918 pandemic influenza outbreak at Camp Custer, MI, and (c) anthrax.

**Plague Epidemic**: The plague epidemics are simulated using a SEIR model with transmission of the disease occurring over a social network. The details of the networked disease model are in [48]. We select a set of index cases depending upon their position at a given moment; thereafter the disease proceeds per the effective reproductive number of the disease that varies in time, as described in [49]. The evolution of an epidemic depends on the individuals designated as index cases; thus one may obtain many different realizations of the epidemic for the same number of

index cases, by varying the individuals (alternatively, by varying the attack site). We designed the time varying reproduction rate such that the epidemic ultimately comes to an end i.e., the final size of the epidemic is a finite number. Figure 3.1(a) shows the progression of the plague epidemic.

**Influenza Epidemic**: The data for the influenza epidemic is obtained from [4]. These are observations of symptomatic patients seeking care at the infirmaries of Camp Custer, MI, on a daily basis, during the 1918 influenza pandemic. Note that the data is not biosurveillance data i.e., there is no visit delay in the observations. Fig. 3.1(b) plots the evolution of the Camp Custer outbreak.

**Anthrax attacks**: The procedure for simulating attacks is fully described in the Appendix of [48]. We consider a population distributed unevenly in space in a square domain. An aerosolized preparation of anthrax is released from the origin (lower left corner) of the domain. The release is evolved in time using a simple Gaussian plume model to provide a time-resolved value of the aerosol concentration at ground level. A breathing rate of 30 l/min is assumed, which is then used to calculate the time-integrated dosage for all the individuals in the population, and using Glassman's formula [21], the probability of infection. The infected individuals are allocated their dose-dependent incubation period (a random variable) per Wilkening's A2 model [58] and a visit delay per the log-normal distribution in [26]. These together determine the time-series of patients who would seek care over a period of time, and serve as the epidemic signature. Figure 3.1(c) shows the progression of the anthrax epidemic.

## 3.4 Results

The inference results for plague, influenza, and anthrax are presented in this section, along with a discussion on the computational expense for each set of tests.

### 3.4.1 Plague Epidemic

We simulate a plague epidemic using the method described in Sec. 3.3. 1000 index cases are infected and the epidemic lasts for 50 days. The epidemic grew to about 15,000 symptomatic cases. The inference was performed using the method described in Sec. 3.1. The inference of plague parameters required $3 \times 10^5 - 5 \times 10^5$ MCMC samples to obtain fully converged statistics. The priors for the model parameters $\Theta = (N_{tot}, \alpha, \tau, \theta_{inc}, \theta_{vd}, \theta_{ir})$ are given in Table 3.1. For certain parameters that are constrained to be positive (or negative), we perform the inversion with their log-transformed values. We generally use Gaussian priors for all parameters except $\alpha$ for which we use a uniform distribution. We found out that the models are sensitive to this parameter, and the inverse problems can generate unphysical solutions with a very small number of secondary cases for highly contagious diseases, unless we impose strict bounds. For the other parameters, the prior standard deviations were set large enough to limit the prior distribution effect on the posterior distributions. A similar approach was taken for other disease models presented in this paper.

We perform the parameter estimation (epidemic characterization) starting 4 or 6 days past the

**Table 3.1.** Prior distributions for the plague model parameters.

| Parameter | Prior distribution |
|---|---|
| $\log(N_{tot})$ | N($\log(10^4)$,2) |
| $\alpha$ | U(0.6,0.99) |
| $\log(-\tau)$ | N($\log(5)$,$\log(10)$) |
| $\log(\theta_{vd})$ | N($\log(0.2)$,1) |
| $\log(\theta_{ir})$ | N($\log(0.1)$,1) |

infection of the index cases. This delay encompasses the time required for the early cases of the disease to develop symptoms and seek care in sufficient numbers and may be thought of as the "alarm" date. The inferences use 3-15 days of data past the date of alarm. The median values, 25th and 75th percentile for a number of parameters entering the plague model are shown in Fig. 3.2. The data is collected starting 4 (blue lines) and 6 (red lines) after the start of the epidemic. For the number of index cases, the true value of around 1000 is bracketed between 25th and 75th percentile using around 7-9 days of data. We also notice that range between the 25th and 75th percentile, approx. 800, remains somewhat independent of the number of days of data used in the inference. The inferred values for the start of the epidemic are shown in Fig. 3.2(b). The true values are -4 and -6 days respectively from the alarm date. In both cases the model overpredicts the magnitude of these values. The total number of symptomatic cases is shown in Fig. 3.2(c). The true value of 15000 is bracketed using 7-9 days of data. Similar to the number of index cases, the 25-75th quantile range remains nearly constant with the number of days of data used to infer the model parameters. The convergence of the Markov chains were monitored using the `mcgibbsit` package [56] in R [47] and these results are independent of the number of samples drawn by the MCMC.

Figure 3.3 shows posterior predictive tests based on the MCMC samples of the plague model parameters. The ensemble of evolutions, based on the MCMC parameter samples, is then used to estimate the median, 25th, and 75th quantile and compare with data series of reported counts. In Fig. 3.3(a), the reported counts from days 5 through 14 (counting from the start of the epidemic) were used to infer the model parameters and then predict the future number of people seeking care, while for Fig. 3.3(b) the results are based on 6 more days of data. In both cases the original data generally lies inside the 25-75th percentile band. In the second case, in Fig. 3.3(b), inclusion of more data in the inference narrows the uncertainty in the expected number of counts compared to Fig. 3.3(a).

### 3.4.2 Influenza Epidemic

Unlike the plague epidemic for which the calculations were based on synthetic data, the computational tests for the influenza epidemic are based on the data collected in Camp Custer, MI, during the 1918 pandemic. The data was obtained from [4]. There is no "canonical" start date for the Camp Custer outbreak. About 10,500 people were affected. The inversion was performed using

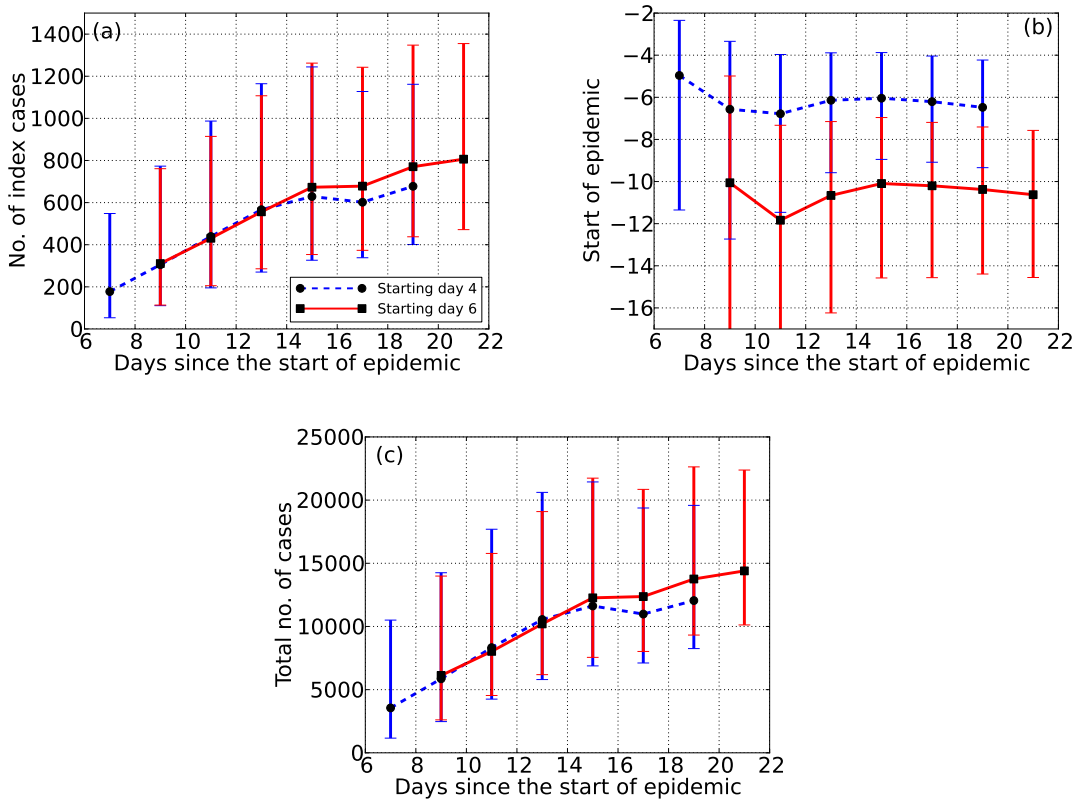


**Figure 3.2.** Estimates for (a) the number of index cases $N_i$, (b) start of the epidemic, $\tau$ and (c) total number of cases for a plague epidemic, $N_{tot}$ for the synthetic plague epidemic. The error bars correspond to the 25th and 75th percentiles, respectively. The blue lines correspond to an alarm date of Day 4 whereas red lines correspond to an alarm date of Day 6.

the method in Sec. 3.1 and $O(10^5)$ MCMC samples were required (similar to plague). The prior distribution for the influenza model parameters are provided in Table 3.2.

Figure 3.4 shows estimates of the start of epidemic and total number of cases, using between 5 and 13 days of data. In the figure, the origin of the horizontal axis is arbitrarily set at the start of data collection, since we do not have a fixed day for the start of the epidemic. The results in Fig. 3.4(a) indicate that the epidemic started approximately 6-8 days before the origin of the horizontal axis. Total number of people who were infected with influenza, approximately 10500, lies within the 25th to 75th percentile band for the results shown in Fig. 3.4(b). The uncertainty in the total number of cases decreases significantly when using up to 11 days of data. Beyond this point the inherent noise in the observations, seen in Fig. 3.1(b), prevent a further decrease in the uncertainty bounds.

Posterior predictive tests for the influenza epidemic are shown in Fig. 3.5. The disease progression

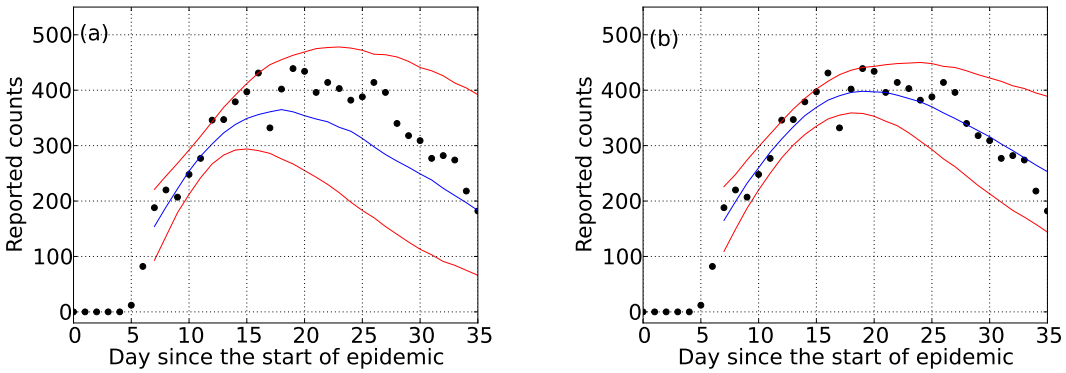


**Figure 3.3.** Posterior predictive tests of the plague epidemic using the MCMC samples. The red lines show the 25th and 75th percentile respectively, while the blue lines show the median. The original data is shown with black circles. The posterior values are based on 9 and 15 days of data in subfigures (a) and (b) respectively. The alarm date is Day 6, and so the start date for the predicted evolutions is Day 7.

**Table 3.2.** Prior distributions for the influenza model parameters.

| Parameter | Prior distribution |
|---|---|
| $\log(N_{tot})$ | N(log(1.2$^4 \times 10^4$),2) |
| $\alpha$ | U(0.95,0.99) |
| $\log(-\tau)$ | N(log(5),log(10)) |
| $\log(\theta_{vd})$ | N(log(12),1) |
| $\log(\theta_{ir})$ | N(log(1.25),0.16) |

estimated based on 9 days of data, in Fig. 3.5(a), shows significant uncertainties beyond Day 10. This result is somewhat expected, given the 5000-15000 uncertainty in the total number of cases, shown in Fig. 3.4(b). The uncertainty range reduces significantly when more data points are included in the computations. Figure 3.5(b) shows posterior predictive tests based on samples computed using 13 days of data.

### 3.4.3 Anthrax Epidemic

We simulated an anthrax outbreak using the method in Sec. 3.3. 22,500 index cases were infected and the epidemic was simulated for 20 days. The inversion was performed as described in Sec. 3.1; note that anthrax is a non-communicable disease and the model being inverted is Eq. 3.1. Figure 3.6 shows statistical results for the number of index cases ($N_i$), start date of the epidemic ($\tau$), and the

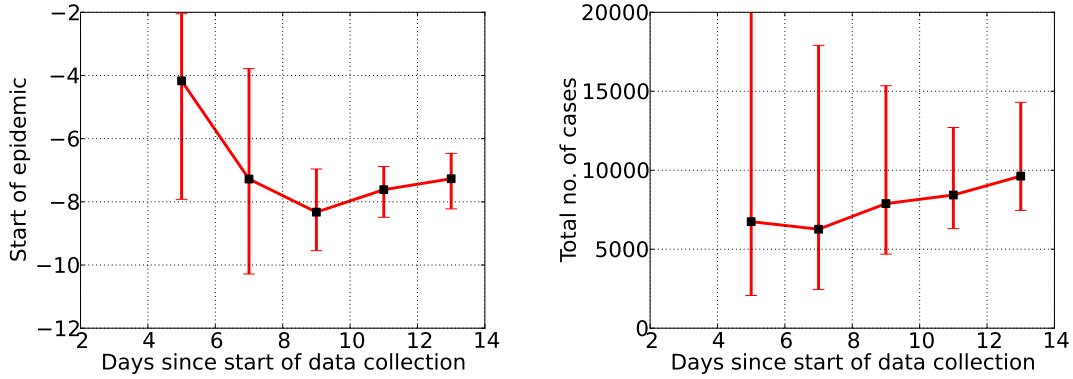


**Figure 3.4.** Estimates for the start of the epidemic (left), and total number of cases (right), for the Camp Custer outbreak. The length of the error bars correspond to the 25th and 75th percentile, respectively.

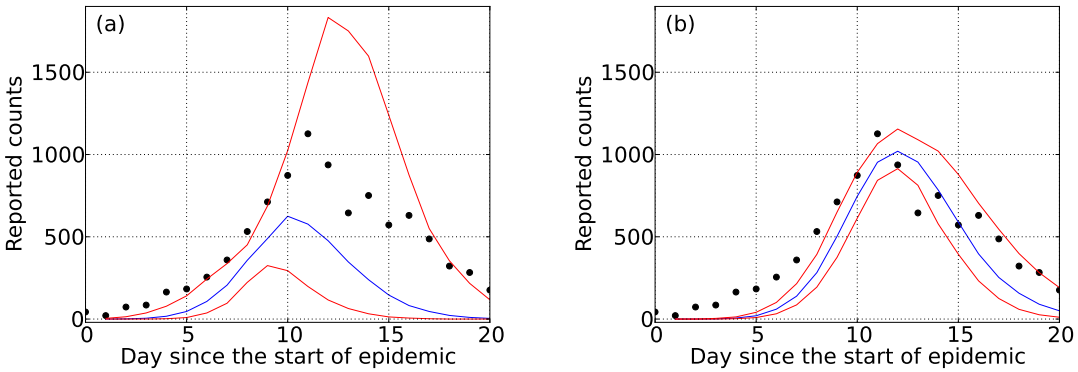


**Figure 3.5.** Posterior predictive tests of the Camp Custer influenza epidemic using the MCMC samples. The red lines show the 25th and 75th percentile respectively, while the blue lines show the median. The original data is shown with black circles. The posterior values are based on 9 and 13 days of data, for subfigures (a) and (b) respectively.

average dose of anthrax spores $D$, as a function of the number of data points used in the inference. The prior distributions for the influenza model parameters are provided in Table 3.3.

**Table 3.3.** Prior distributions for the anthrax model parameters.

| Parameter | Prior distribution |
|---|---|
| $\log(N_i)$ | N($10^4$,10) |
| $\log(-\tau)$ | N(0,log(10)) |
| $\log_{10} D$ | N(3,2) |
| $\log(\theta_{vd})$ | N(0,1) |

The true value for $N_i$ in Fig. 3.6(a), $22,500$, is reasonably well estimated within 8-10 days after the start of the epidemic, i.e., by using an observation period 4 to 6 days long. Note that the inferences initially show large uncertainties (the error-bars indicate the inter-quartile range) but decrease as more data become available. The inferred time of the start of the epidemic is shown in Fig. 3.6(b). The results are calculated with respect with to the start of data collection, 4 and 6 days after the start of the actual epidemic. The results are within 1 day for both cases and the difference between 25th and 75th quantiles decreases to about 1 day which is the resolution of data collection. The average dose $\log_{10} D = 2.8$ is bracketed within 25-75th quantiles using 3-5 days of data. The median value agreement with the actual value improves with the number of data points used to infer the parameter, however the uncertainty does not reduce much beyond 7 days of data.

Figure 3.7 shows posterior predictive tests constructed in a similar fashion as were the results for plague and influenza tests. In Fig. 3.7(a), the reported counts from Days 5 through 10 were used to infer the model parameters and then predict the future numbers of sick people requesting care. In Fig. 3.7(b), the results are based on 4 more days of data, from Day 5 through Day 14. Due to the additional information contained in these data points, the uncertainty in the reported counts is smaller compared to the results in Fig. 3.7(a).

### 3.4.4 Computational Expense

For the plague and influenza computations, the models for the number of sick people seeking care on a daily basis require the evaluation of the single and double integrals in (Eq. 3.1) and (Eq. 3.2) corresponding to the number of index and secondary cases, respectively. The evaluation of the double integral is expensive. For the anthrax computations, only the number of index cases are computed since the disease is not contagious. The computational times presented in Table 3.4 are for runs on a 2.6GHz Intel Core 2 Duo.

The CPU times correspond to a full set of parameter inferences, e.g. using between 5-15 days of data for plague. The evaluation of the double integral in Eq. (3.2) significantly increases the computational cost for the estimation of plague and influenza model parameters. For these computations the CPU time is one order of magnitude larger compared to the anthrax. A surrogate

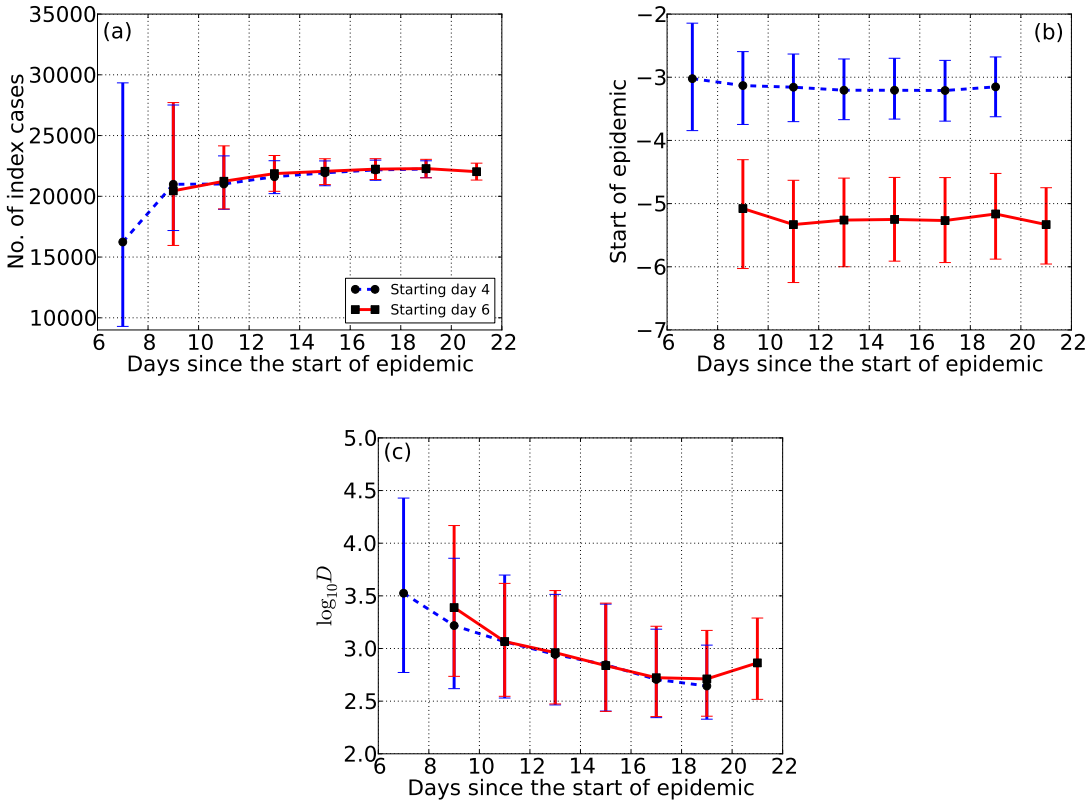


**Figure 3.6.** Estimates for (a) the number of index cases, (b) start of the epidemic, and (c) dose magnitude for the anthrax epidemic. The length of the error bars correspond to the 25th and 75th quantiles, respectively. The blue and red lines correspond to alarm dates of Day 4 and 6 respectively.

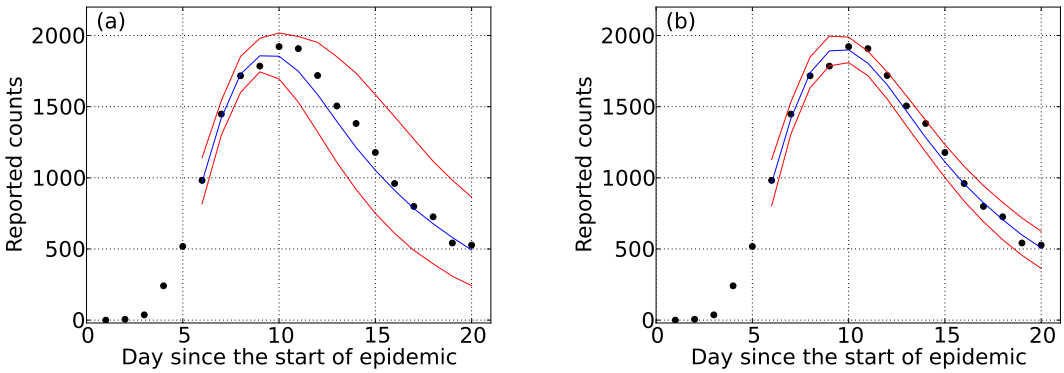


**Figure 3.7.** Posterior predictive tests of the anthrax outbreak progression using the MCMC samples. The red lines show the 10th and 90th percentile respectively, while the blue lines shows the median. The original data is shown with circles. The posterior values are based on 5 and 9 days of data, respectively. The alarm date is Day 6, and so the start date for the predicted evolutions is Day 7.

**Table 3.4.** Computational expense for the inference of plague, influenza, and anthrax parameters.

| Model | Cases | No. of samples | Time-series length [days] | CPU time [h] |
|---|---|---|---|---|
| Plague | Index & Sec | $3 \times 10^5$ | 5-15 | 7.0 |
| Influenza | Index & Sec | $3 \times 10^5$ | 5-13 | 5.6 |
| Anthrax | Index | $5 \times 10^5$ | 3-15 | 0.2 |

model approach is introduced in the next section, in order to reduce the computational expense of the plague and influenza models.

# Chapter 4

# Surrogate Models

The surrogate model approach consists of replacing expensive models with polynomial functions, which are far cheaper to compute (versus the full epidemic model) but are accurate enough compared to the full model evaluation. Specifically, we will use polynomial chaos (PC) expansions [19, 59] to construct a surrogate model for the number of symptomatic people as a function of time.

## 4.1 Construction of Surrogate Models

Let $f(x)$ be an expensive model that depends on an array of random variables $x = (x_1, x_2, \ldots, x_n)$. For example, for the plague model $x = \Theta_P = (t, N_{tot}, \alpha, \tau, \theta_{ir}, \theta_{vd})$. We will approximate the model $f$ as a polynomial expansion:

$$f(x) \approx \sum_{k=0}^{P} c_k \Psi_k^{(n)}(x), \tag{4.1}$$

where $\Psi_k^{(n)}(x) = \Psi_{k_1}(x_1) \cdot \Psi_{k_2}(x_2) \cdot \ldots \cdot \Psi_{k_n}(x_n)$ are multi-variate polynomials obtained by taking the product of uni-variate polynomials that are functions of each component $x_i$ in the array of random variables $x$ [29]. Typically, these polynomials form an orthogonal basis in order to minimize the numbers of terms $P$ required to obtain certain accuracy in the approximation. Henceforth we will drop the superscript "(n)" to simplify the notation. The multi-variate polynomials $\Psi_k(x)$, can be chosen to be orthogonal with respect to the probability density function of $x$ ($g(x)$) in order to obtain surrogate models that are most accurate where $x$ is most likely.

$$\int_D \Psi_k(x)\Psi_l(x)g(x)dx = \delta_{kl}\alpha_k \tag{4.2}$$

The expansion coefficients $c_k$ (in Eq. 4.1) can be computed in a number of ways; we use the Galerkin approach that exploits the orthogonality of the terms in the expansion

$$c_k = \frac{\langle f(x)\Psi_k(x)\rangle}{\langle \Psi_k(x)\Psi_k(x)\rangle}, \quad \text{and} \quad \langle f(x)\Psi_k(x)\rangle = \int_D f(x)\Psi_k(x)g(x)dx \tag{4.3}$$

The integrals necessary to evaluate $\langle f(x)\Psi_k(x)\rangle$ are evaluated using numerical quadrature

$$\langle f(x)\Psi_k(x)\rangle \approx \sum_{q=1}^{N_q} f(x_q)\Psi_k(x_q)w_q. \tag{4.4}$$

Here, $x_q$ and $w_q$ are the quadrature points and weights, respectively, for the quadrature formula used to compute the integral. Therefore, in order to evaluate the coefficients of the PC expansion, the full model $f$ needs to be evaluated at specific values of the input parameters $x$, chosen to correspond to the quadrature points needed for the numerical evaluation of the projection formulas. In this paper we will be using Legendre polynomials, orthogonal with respect to uniform probability distributions. In order to construct expansions based on Legendre polynomials the parameter ranges will be rescaled to $[-1,1]$ intervals.

Once a PC approximation for the epidemiological model is fully constructed, this approximation can replace the evaluations of the full model in the MCMC procedure described above. The MCMC technique proceeds as usual to determine the distribution of model parameters that best fits the epidemiological data.

## 4.2 Surrogate Models for Plague

Careful examination of the equations for the number of index cases (3.1) and secondary cases (3.2) reveals that the daily counts of the number of people becoming sick can be written as:

$$\nu_{tot}(t) = N_{tot}(\alpha f_1(t-\tau,\theta_{vd}) + (1-\alpha) f_2(t-\tau,\theta_{vd},\theta_{ir})) \tag{4.5}$$

Here $f_1$ is the integral in Eq. (3.1), while $f_2$ is the double integral in Eq. (3.2). In this form, parameters $N_{tot}$ and $\alpha$ are proportionality factors, while $\tau$ leads to a shift in the disease evolution depending on the start date. This allows us to reduce the number of dimensions from six to three when writing the polynomial expansion (4.1):

$$\nu_{tot}(t) = N_{tot} \sum_{k=1}^{P} (\alpha c_{1,k} + (1-\alpha) c_{2,k}) \Psi_k(t-\tau,\theta_{vd},\theta_{ir})) \tag{4.6}$$

The superscript $(3)$ in Eq. (4.6) indicates that $\Psi_k$'s are trivariate polynomials. The domain of integration $D$ in Eq. 4.3, used to calculate $c_{i,k}$ is $0 \le t \le 50, 10^{-2} \le \theta_{vd} \le 2, 10^{-2} \le \theta_{ir} \le 2$. This domain was chosen large enough to ensure the surrogate model is accurate over the entire range of parameters that can be encountered during the inversion process.

Figure 4.1 shows the evolution of the number of people seeking care for a range of $\theta_{ir}$ values. In this figure, $N_{tot}$, $\alpha$, and $\theta_{vd}$ are set to $10^4$, 0.92, and 0.3, respectively. Several polynomial orders, from 5 through 19, are considered. Visual inspection shows the agreement between the full model, in blue, and the surrogate models, in red, is quite bad for polynomials of order 5 and 11, but steadily improves as the order increases to 19. For the surrogate models using 19th order polynomials, approximately 20000 model evaluations were necessary to compute the PC coefficients.

Figure 4.2(a) shows epidemic curves corresponding to several slices through the surfaces shown in Figure 4.1(d). The $\theta_{ir}$ values are shown near each set of curves in Fig. 4.2(a) and the color scheme is the same as in the previous figure. Some discrepancies are observed between the full and surrogate models corresponding to small $\theta_{ir}$. These discrepancies are inherent to polynomial

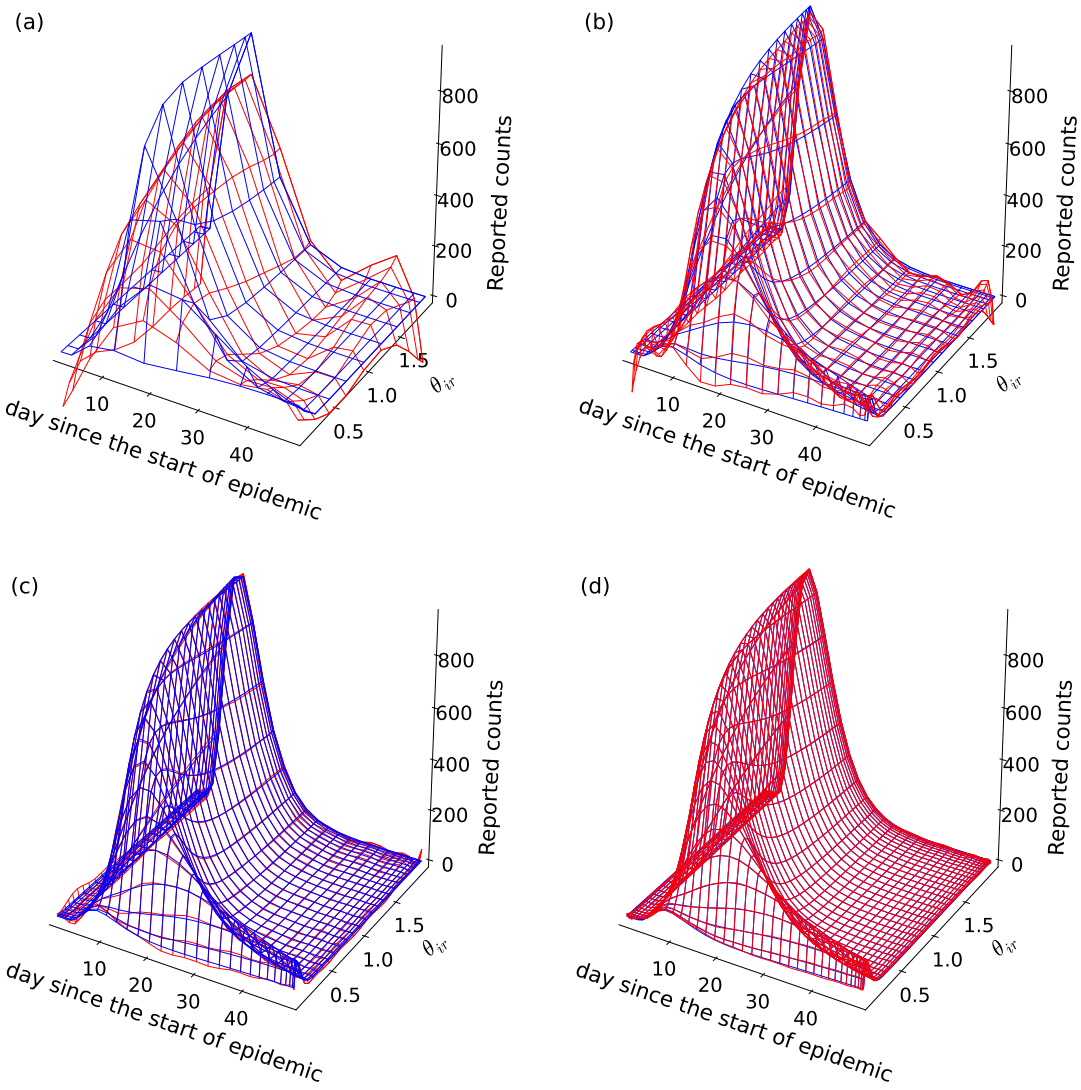


**Figure 4.1.** Evolution of plague epidemic as a function of the infection rate parameter. The blue wireframe results are based on the full model, while the red wireframes correspond to surrogate models using 5, 11, 15, and 19 order polynomials. All results correspond to $\theta_{vd} = 0.3$. Subfigures (a) and (b) show poor comparisons but the 19th order polynomial in (d) shows good agreement.

approximations for highly non-linear functions. To circumvent this problem we introduce an alternate representation for the surrogate model, where the dependency on the $\theta_{vd}$ and $\theta_{ir}$ is replaced with a dependency on the natural logarithms of these parameters:

$$\nu_{tot}(t) = N_{tot} \sum_{k=1}^{P} (\alpha c_{1,k} + (1-\alpha) c_{2,k}) \Psi_k(t-\tau, \log(\theta_{vd}), \log(\theta_{ir}))) \tag{4.7}$$

This modification naturally adds weight to the lower range of values for both $\theta_{vd}$ and $\theta_{ir}$. Figure 4.2(b) shows epidemic curves from surrogate models constructed using Eq. (4.7). The new results show a better agreement for epidemic curves corresponding to small values of $\theta_{ir}$. Similar results are also obtained for the range of $\theta_{vd}$ values.

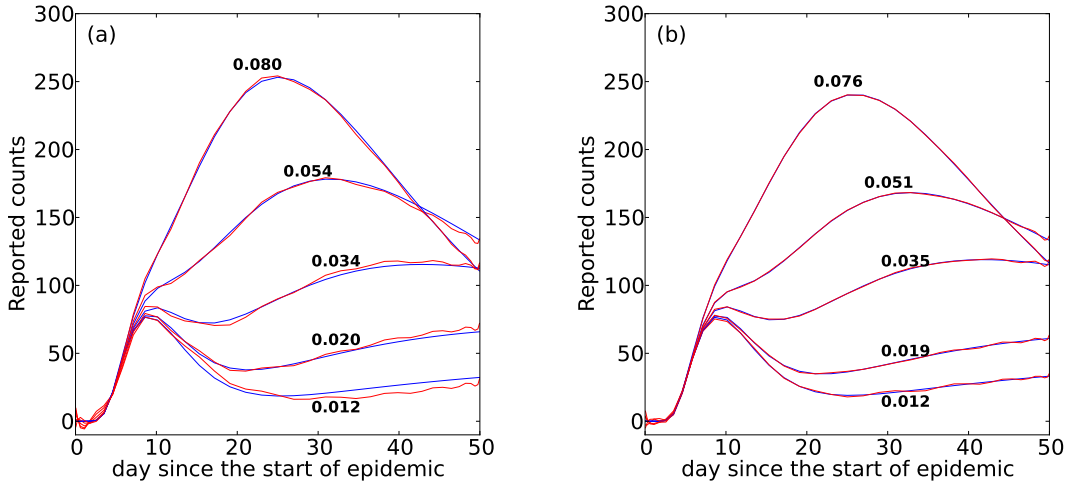


**Figure 4.2.** Evolution of plague epidemic for several infection rate parameter values. Left frame surrogate models are based on Eq. (4.6) while those in the right frame are based on Eq. (4.7). The surrogate models use 19th order polynomials.

We repeated the epidemic characterization runs described in Sec. 3.4.1 using the surrogate model in Eq. (4.6). The results in Fig. 4.3 show the inferred $N_{tot}$ values using the alternative approximations Eq. (4.6) and (4.7). The later surrogate model formulation which exhibits a better agreement with the full model at smaller $\theta_{vd}$ and $\theta_{ir}$ values also does a better job estimating the $N_{tot}$ range of values. While in Fig. 4.3(a) the results show little convergence with polynomial order, in Fig. 4.3(b) there is a clear improvement when using polynomial expansions based on the transformed parameters in Eq. (4.7). Similar agreement is observed for the other parameters comprising the epidemic model for plague.

## 4.3 Surrogate Models for Influenza

We attempted to use the same surrogate modeling approach as described in Sec. 4.2 using Eq. (4.7), but applied to an influenza model. We found it impossible to identify a single polynomial order that could represent $\nu_{tot}(t)$ accurately in the entire domain ($0 \leq t \leq 50, 5 \leq \theta_{vd} \leq 25, 0.9 \leq \theta_{ir} \leq 1.6$. Consequently, we partitioned the $t$ dimension and fit a separate surrogate model in each. Partitioning the $t$-dimension resulted in regions where the full model behaved very smoothly and

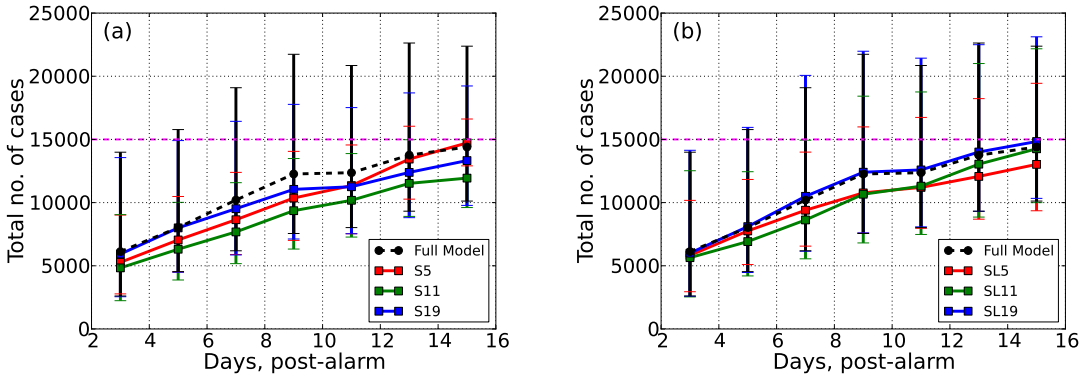


**Figure 4.3.** Estimates for total number of cases for a plague epidemic. The length of the error bars correspond to the 25th and 75th percentiles, respectively. The left frame surrogate model results correspond to Eq. (4.6), using 5th (S5), 11th (S11), and 19th (S19) order polynomials. The left frame results correspond to Eq. (4.7) and the same sequence of polynomial fits. The dashed line shows the actual $N_{tot}$ value.

could be fit with relatively low-order polynomials e.g. order 9. The surrogate model, thus consists of a collection of polynomial approximations, each valid in its own partition of the parameter space. The partitioning approach combined with low order polynomials can reduce the total number of model evaluations. For a collection of 10 partitions with 9th order PC expansions in each partition, about 13000 model evaluations (compared to 20000 in the previous section) were necessary to compute the PC coefficients.

In Figs. 4.4 and 4.5, we explore the impact of the polynomial expansion order on the quality of the surrogate models, as well as the effect of increasing the number of partitions of the $t$ dimension. We see that, due to smooth model behavior in each domain partition, 5th to 9th order polynomials are sufficient to capture the full model behavior for 5-day partitions (in Fig. 4.5).

We then repeat the epidemic characterizations in Sec. 3.4.2 but with the original epidemic model replaced with its surrogate. The results in Figs. 4.6 and 4.7 show the effects of the partition size and polynomial expansion order, respectively, on the inferred 25th percentile, 75th percentile, and median values for the total number of cases and infection rate parameters. Both sets of results show a clear improvement in the accuracy of results when reducing the partition size and/or increasing the polynomial order. From Figs. 4.6 and 4.7, we find that 10-day partitions with 9th order polynomials may provide surrogate models sufficiently accurate for use in epidemic characterizations.

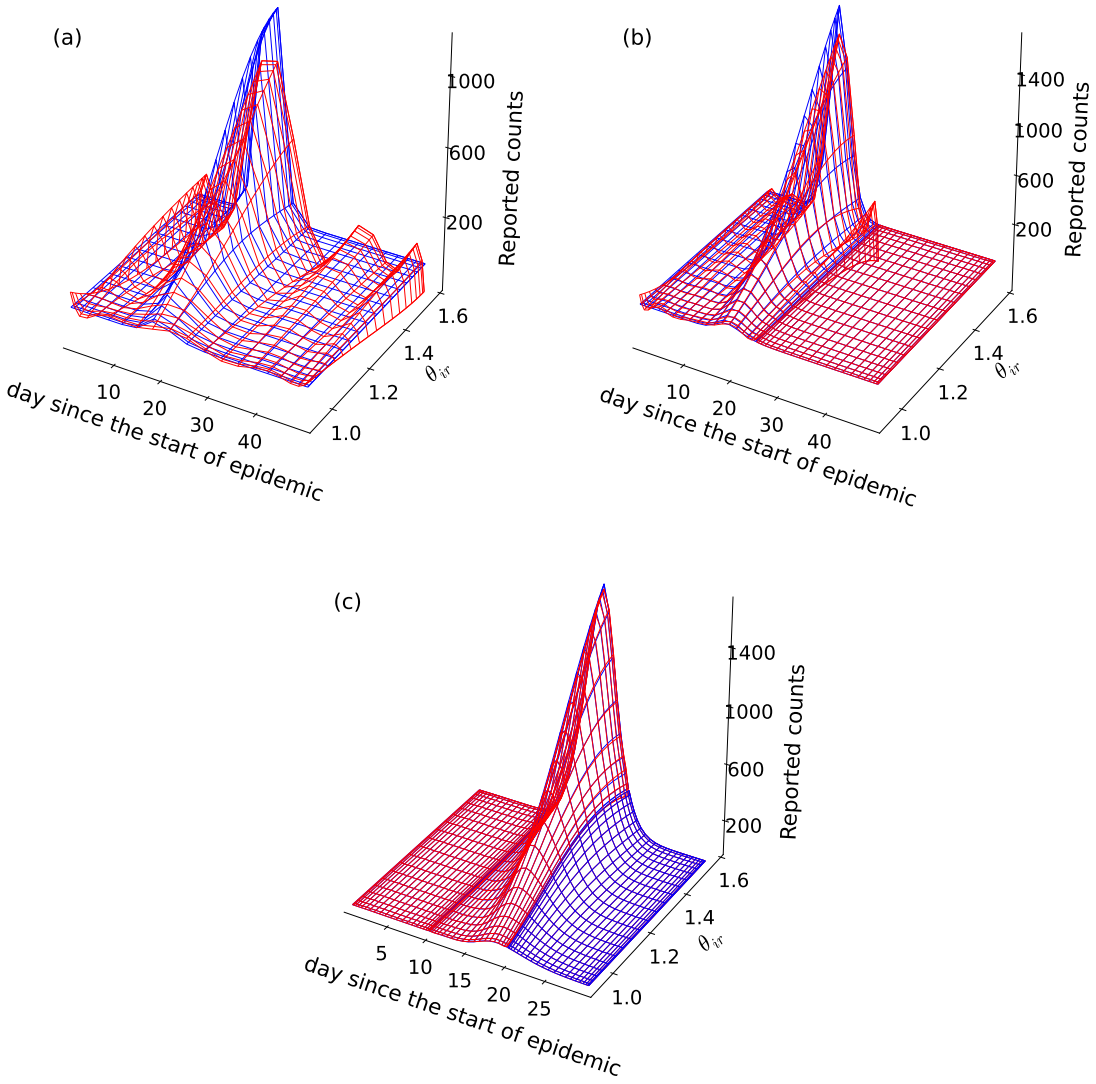


**Figure 4.4.** Evolution of influenza epidemic as a function of the infection rate parameter. The full model is shown in blue and the surrogate models in red. The day axis consists of (a) one 50-day partition, (b) two 25-day partitions, and (c) five 10-day partitions. All surrogate models use 9th order polynomial expansions and the results correspond to $N_{tot} = 10^4$, $\alpha = 0.99$, and $\theta_{vd} = 0.3$.

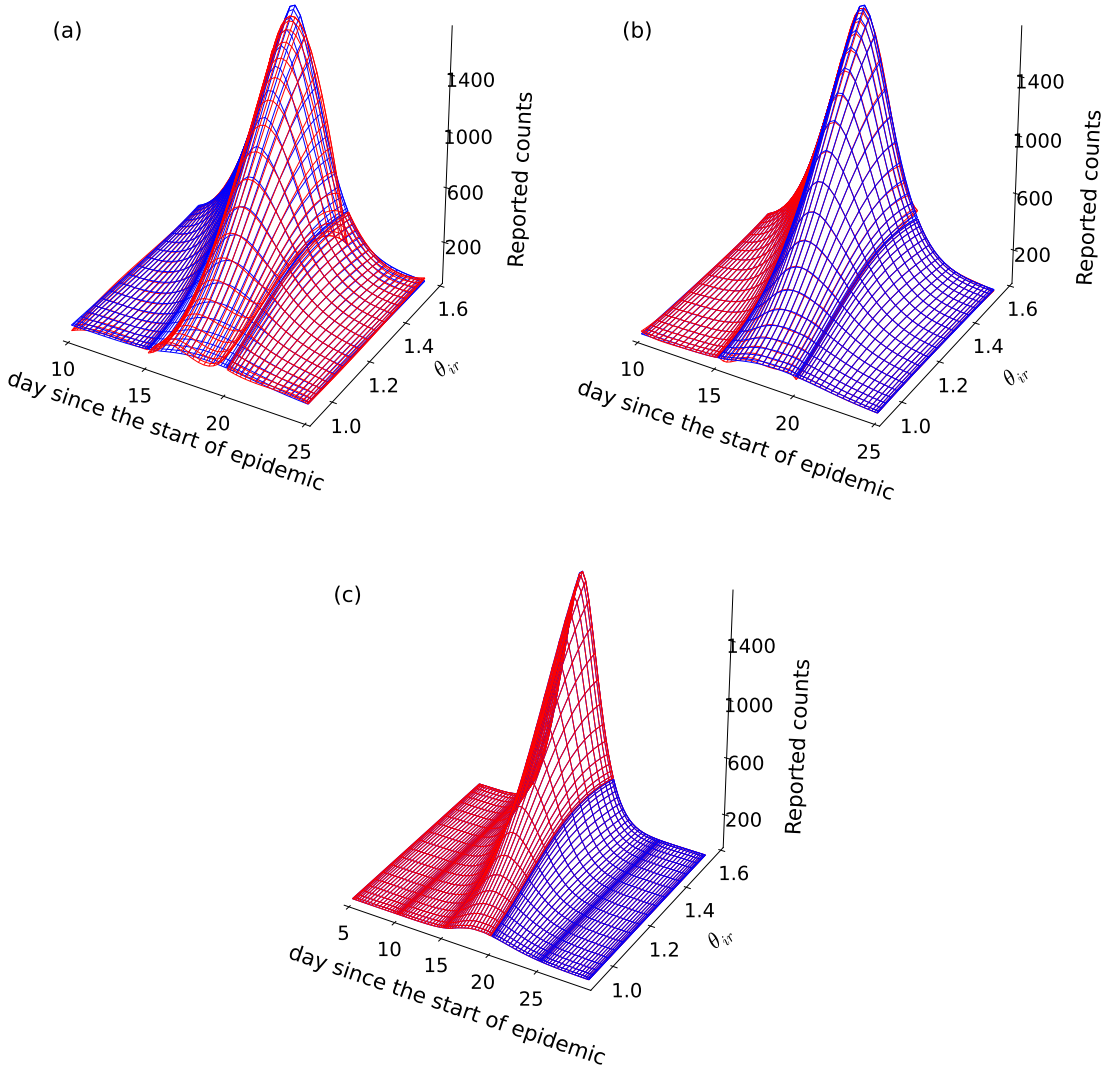


**Figure 4.5.** Evolution of influenza epidemic as a function of the infection rate parameter. The full model is shown in blue and the surrogate models in red. The surrogate models use (a) 3-rd order, (b) 5th order, and (c) 9th order polynomial expansions. For all frames the day axis consists of 10 5-day partitions. The values for the other parameters are the same as for Fig. 4.4.

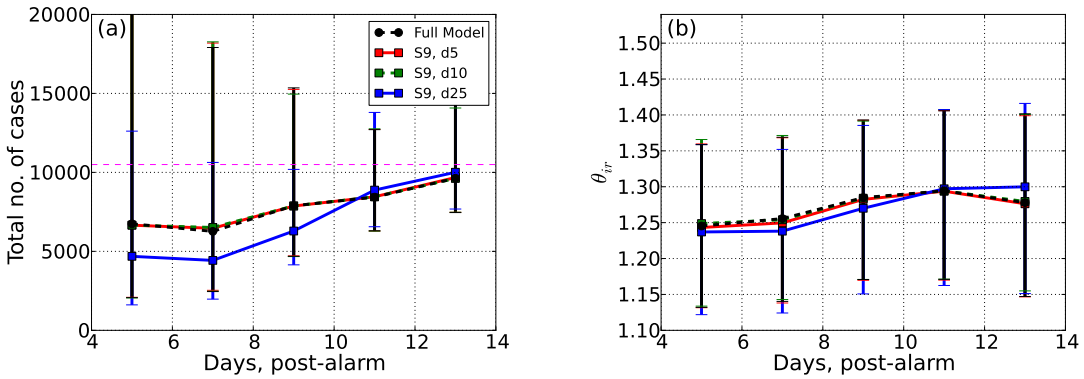


**Figure 4.6.** Influenza model parameters, $N_{tot}$ and $\theta_{ir}$, estimated using 9th order polynomials, split over 5-day (d5), 10-day (d10), and 25-day (d25) intervals.

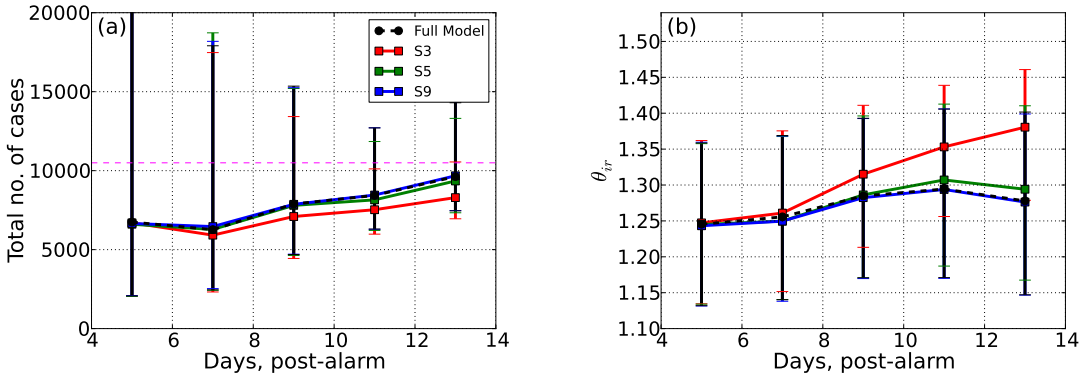


**Figure 4.7.** Influenza model parameters, $N_{tot}$ and $\theta_{ir}$, estimated using 3rd (S3), 5th (S5), and 9th (S9) order polynomials split over 5 day intervals.

## 4.4 Computational Expense

The computational time required for the surrogate model approach in Table 4.1 show an almost two orders of magnitude speed-up on the same computing platform compared to the results based on the direct plague and influenza model evaluations in Table 3.4. The CPU time values shown here do not include the CPU times required to generate the coefficients for the polynomial expansions. These values depend on the polynomial order and the partition size, and are between 0.05-0.1 CPU hours. However, once computed, the polynomial expansions can be reused to infer the epidemic model parameters for several data sets. This amortization renders their computational cost negligible.

**Table 4.1.** Computational expense for the inference of plague, influenza, and anthrax parameters using the surrogate model approach.

| Model | No. of samples | Data [days] | CPU time [h] |
|---|---|---|---|
| Plague | $3 \times 10^5$ | 5-15 | 0.14 |
| Influenza | $3 \times 10^5$ | 5-13 | 0.11 |

This page intentionally left blank.

# Chapter 5

# Summary and Conclusions

This report presents an approach for the statistical characterization of partially observed epidemics using surrogate models. Data consists of time series of symptomatic patients diagnosed with the disease. The characterization is performed using an epidemic model, which contains submodels for the incubation period, visit delay, and infection rat. The submodels are specialized for three different diseases (anthrax, plague and influenza). The total number of cases, start of epidemic, and other epidemiological parameters are estimated from the available time series using a deconvolution approach. The characterization problem is formulated as a Bayesian inverse problem, and epidemiological parameters are estimated as distributions using a Markov chain Monte Carlo (MCMC) method, thus quantifying the uncertainty in the estimates.

We find that epidemiological models that have the ability to reproduce the complex temporal dynamics of epidemics (generally those of communicable diseases) cannot be naively used in "real-time" characterization studies with MCMC. Scalable techniques like Ensemble Kalman Filters/Smoothers may allow their use, but only if Gaussian assumptions are made regarding the distribution of the estimated parameters. This is best avoided within the context of sparse data. We introduce a competing approach, where the epidemiological model is replaced by its surrogate. The surrogate model is a polynomial expansion created by projecting the output of the epidemiological model on a set of orthogonal polynomial bases; thereafter, computations involving the surrogate model reduce to evaluations of a polynomial. We achieve more than a factor of 10 speed-up when we do so, with little or no loss of accuracy. We find that the number of sample points at which the epidemic model has to be evaluated prior to projection is $\mathrm{O}(10) - \mathrm{O}(10^2)$ fewer than the number of samples required by MCMC to converge; thus it may not even be necessary to construct the surrogate models offline. This advantage arises partially due to our choice of the basis set (polynomial chaos) and partly due to the large number of MCMC samples required to explore the parameter space. These results were obtained using synthetic epidemic data for anthrax and plague outbreaks, and data from the 1918 influenza pandemic collected at Camp Custer, Michigan.

We could not find a systematic way of constructing the surrogate model. In one case, the surrogate model consisted of high-order ($19^{th}$-order) functions of the log-transformed input parameters, whereas in the other, the parameter domain had to be partitioned and fitted with relatively lower-order polynomials. The particular approach adopted is dependent on the behavior of the model in question as well as the region in the parameter space where accuracy is desired. While we adopted domain partitioning and stretching, the same could potentially be accomplished by sampling the parameter domain in an uneven or adaptive manner, predicated on the model response (or

its gradient).

Biosurveillance networks are becoming ubiquitous and are increasingly used to detect the start of outbreaks. As the accuracy and timeliness of their data improves (and the quantity increases), automated processing, with a view of detecting patterns or drawing inferences, will gain epidemiological and public health relevance. Accelerated means of doing so, along with a quantification of uncertainty in the inferences, can be expected to assume practical importance. In this paper, we have demonstrated an approach to do so, without significant loss of accuracy. While the use of (polynomial chaos) surrogate models may be novel in epidemiology, they are nevertheless used widely in design and optimization efforts in other fields. Consequently, they may potentially be useful in real-time epidemiology too.

# References


[1] Centers for Disease Control and Prevention, Seasonal Influenza (Flu), Past Weekly Surveillance Reports, February 2011.

[2] L. M. A Bettencourt. An ensemble trajectory method for real-time modeling and prediction of unfolding epidemic: Analysis of the 2005 Marburg fever outbreak in Angola. In G. Chowell et al, editor, *Mathematical and Statistical Estimation Approaches in Epidemiology*, pages 143–161. Springer Science+Business Media B.V., 2009.

[3] L. M. A. Bettencourth and R. M. Ribeiro. Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *Public Library of Science*, 3(5), 2008. e2185.

[4] J. N. Bombardt and H. E. Brown. Potential influenza effects on military populations. Technical Report paper P-3786, Institute for Defense Analysis, 2003.

[5] T. Britton and P. ONeill. Bayesian inference for stochastic epidemics in populations with random social structure. *Scandinavian Journal of Statistics*, 29:375–390, 2002.

[6] R. Brookmeyer, N. Blades, M. Hugh-Jones, and D. A. Henderson. The statistical analysis of truncated data: application to the Sverdlovsk anthrax outbreak. *Biostatistics*, 2:233–247, 2001.

[7] R. Brookmeyer and M. H. Gail. A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association*, 83:301–308, 1988.

[8] S. Cauchemez, P.-Y. Boelle, C. A. Donnelly, N. Ferguson, G. Thomas, G. M. Leung, A. J. Hedley, R. M. Anderson, and A.-J. Valleron. Real-time estimates in early detection of SARS. *Emerging Infectious Diseases*, 12(1):110–113, 2006.

[9] S. Cauchemez, P.-Y. Boelle, G. Thomas, and A.-J. Valleron. Estimating in real time the efficacy of measures to control emerging communicable diseases. *American Journal of Epidemiology*, 164:591–597, 2006.

[10] S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron, and P.Y. Boelle. A bayesian mcmc approach to study transmission of influenza: Application to household longitudinal data. *Statistics in Medicine*, 23:3469–3487, 2004.

[11] P. G. Constantine, A. Doostan, Q. Wang, and G. Iaccarino. A surrogate accelerated bayesian inverse analysis of the hyshot ii supersonic combustion data. In *Proceedings of the Summer Program 2010, Center for Turbulent Research*, 2010.

[12] B.J. Debusschere, H.N. Najm, A. Matta, O.M. Knio, R.G. Ghanem, and O.P. Le Maître. Protein labeling reactions in electrochemical microchannel flow: Numerical simulation and uncertainty propagation. *Physics of Fluids*, 15(8):2238–2250, 2003.

[13] M. Eichner and K. Dietz. Transmission potential of smallpox: Estimates based on detailed data from an outbreak. *American Journal of Epidemiology*, 158:110–117, 2003.

[14] M. Frangos, Y. M Marzouk, K. Willcox, and B. van Bloemen Waanders. Surrogate and reduced-order modeling: A comparison of approaches for large-scale statistical inverse problems. In L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, Y. Marzouk, L. Tenorio, B. van Bloemen Waanders, and K. Willcox, editors, *Large-Scale Inverse Problems and Quantification of Uncertainty*. John Wiley & Sons, Ltd, Chichester, UK, 2010.

[15] D. Gamerman. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall, London, 1997.

[16] R. Gani and S. Leach. Transmission potential of smallpox in contemporary populations. *Nature*, 414:748–751, 2001.

[17] R. Ghanem. Probabilistic characterization of transport in heterogeneous media. *Computational Methods in Applied Mechanics and Engineering*, 158:199–220, 1998.

[18] R.G. Ghanem, J.R. Red-Horse, and A. Sarkar. Modal properties of a space-frame with localized system uncertainties. In A. Kareem, A. Haldar, B.F. Spencer Jr., and E.A. Johnson, editors, *8th ASCE Specialty Conference of Probabilistic Mechanics and Structural Reliability*, number PMC200-269. ASCE, 2000.

[19] R.G. Ghanem and P.D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer Verlag, New York, 1991.

[20] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 1996.

[21] H. N. Glassman. Industrial inhalational anthrax. *Bacteriological reviews*, 30:657–659, 2007.

[22] M. Kakehashi H. Nishiura, M. Schwehm and M. Eichner. Transmission potential of primary pneumonic plague: time-inhomogeneous evaluation based on historical documents of the transmission network. *Journal of Epidemiology and Community Health*, 60:640–645, 2006.

[23] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7:223–242, 2001.

[24] D. Higdon. A primer on space-time modeling from a Bayesian perspective. In Barbel Finkenstadt, Leonhard Held, and Valerie Isham, editors, *Statistical methods for spatio-temporal systems*. CRC Press, 2007.

[25] W. R. Hogan, G. F. Cooper, G. L. Wallstrom, M. M. Wagner, and J.-M. Depinay. The bayesian aerosol release detector: An algorithm for detecting and characterizing outbreaks caused by an atmospheric release of bacillus anthracis. *Statistics in Medicine*, 26:5225–5252, 2007.

[26] W. R. Hogan and G. L. Wallstrom. Approximating the sum of lognormal distributions to enhance models of inhalational anthrax. In *Quantitative Methods for Defense and National Security*, 2007.

[27] J.-E. C. Holty, D. M. Bravata, H. Liu, R. A. Olshen, K. M. McDonald, and D. K. Owens. Systematic review: A century of inhalational anthrax cases from 1900 to 2005. *Annals of Internal Medicine*, 144:270–280, 2006.

[28] O.P. Le Maître, R.G. Ghanem, O.M. Knio, and H.N. Najm. Uncertainty propagation using Wiener-Haar expansions. *Journal of Computational Physics*, 197(1):28–57, 2004.

[29] O.P. Le Maître and O.M. Knio. *Spectral Methods for Uncertainty Quantification*. Springer, New York, NY, 2010.

[30] S. Lefantzi and J. Ray. Deriving a model for influenza epidemics from historical data. Technical Report SAND2011-6633, Sandia National Labs, 2011.

[31] J. Legrand, J. R. Egan, I. M. Hall, S. Cauchemez, S. Leach, and N. M. Ferguson. Estimating the location and spatial extent of a covert anthrax release. *PLoS Computational Biology*, 5:e1000356, 2009.

[32] J. Liao and R. Brookmeyer. An empirical Bayes approach to smoothing in backcalculation of HIV infection rates. *Biometrics*, 51:579–588, 1995.

[33] C. Lieberman, K. Willcox, and O. Ghattas. Parameter and state model reduction for large-scale statistical inverse problems. 32(5):2523–2542, 2010.

[34] M. Lipsitch, T. Cohen, B. Cooper, J. M. Robins, S. Ma, L. James, G. Gopalakrishna, S. K. Chew, C. C. Tan, M. H. Samore, D. Fisman, and M. Murray. Transmission dynamics and control of Severe Acute Respiratory Syndrome. *Science*, 300:1966–1970, 2003.

[35] X. Ma and N. Zabaras. An efficient bayesian inference approach to inverse problems based on an adaptive sparse grid collocation method. *Inverse Problems*, 25:035013, 2009.

[36] M. A. Martinez-Beneito, P. Botella-Rocamora, and O. Zurriaga. A kernel-based spatio-temporal surveillance system for monitoring influenza like incidence. *Statistical Methods in Medical Research*, 00:1–16, 2010.

[37] Y. M. Marzouk and H. N. Najm. Dimensionality reduction and polynomial chaos acceleration of bayesian inference in inverse problems. *Journal of Computational Physics*, 228:1862–1902, 2009.

[38] Y. M. Marzouk, H. N. Najm, and L. A. Rahn. Stochastic spectral methods for efficient bayesian solution of inverse problems. *Journal of Computational Physics*, 224:560–586, 2007.

[39] L. Mathelin, C. Desceliers, and M. Hussaini. Stochastic data assimilation of the random shallow water model loads with uncertain experimental measurements. *Computational Mechanics*, 47:603–616, 2011.

[40] M. Meselson, J. Guillemin, M. Hugh-Jones, A. Langmuir, I. Popova, A. Shelokov, and O. Yampolskaya. The Sverdlovsk anthrax outbreak of 1979. *Science*, 266:1202–1208, 1994.

[41] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

[42] H. Nishiura and G. Chowell. The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In G. Chowell et al, editor, *Mathematical and Statistical Estimation Approaches in Epidemiology*, pages 103–121. Springer Science+Business Media B.V., 2009.

[43] H. Nishiura and M. Kakehashi. Real-time estimation of reproduction numbers based on case notifications – Effective reproduction number of primary pneumonic plague. *Tropical Medicine and Health*, 33:127–132, 2005.

[44] A. O'Hagan. Bayesian analysis of computer code outputs: a tutorial. *Reliability Engineering and System Safety*, 91:1290–1300, 2006.

[45] J. Boon Som Ong, M. I-Cheng Chen, A. R. Cook, H. Chyi Lee, V. J. Lee, R. Tzer Pin Lin, P. Ananth Tambyah, and L. gan Goh. Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in singapore. *Public Library of Science One*, 5(4), 2010. e10036.

[46] N. V. Queipo, R. T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, and P. K. Tucker. Surrogate-based analysis and optimization. *Progress in Aerospace Sciences*, 41:1–28, 2005.

[47] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[48] J. Ray, B. M. Adams, K. D. Devine, Y. M. Marzouk, M. M. Wolf, and H. H. Najm. Distributed micro-releases of bioterror pathogens: threat characterizations and epidemiology from uncertain patient observables. Technical Report SAND2008-6044, Sandia National Labs, 2008.

[49] J. Ray, P. T. Boggs, D. M. Gay, M. N. Lemaster, and M. E. Ehlen. Risk-based decision making for staggered bioterrorist attacks: Resource allocation and risk reduction in reload scenarios. Technical Report SAND2009-6008, Sandia National Labs, 2009.

[50] J. Ray, Y. M. Marzouk, and H. N. Najm. A Bayesian approach for estimating bioterror attacks from patient data. *Statistics in Medicine*, 30(2):101–126, 2011.

[51] S. Riley, C. Fraser, C. A. Donnelly, A. C. Ghani, L. J. Abu-Raddad1, A. J. Hedley, G. M. Leung, L.-M. Ho, T.-H. Lam, T. Q. Thach, P. Chau, K.-P. Chan, S.-V. Lo, P.-Y. Leung, T. Tsang, W. Ho, K.-H. Lee, E. M. C. Lau, N. M. Ferguson, and R. M. Anderson. Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. *Science*, 300:1961–1966, 2003.

[52] K. Sargsyan, B. Debusschere, H. Najm, and O. Le Maître. Spectral representation and reduced order modeling of the dynamics of stochastic reaction networks via adaptive data partitioning. *SIAM Journal on Scientific Computing*, 31(6):4395–4421, 2010.

[53] J. Walden and E. H. Kaplan. Estimating the time and size of a bioterror attack. *Emerging Infectious Diseases*, 10:1202–1205, 2004.

[54] J. Wallinga and P. Teunis. Different epidemic curves for Severe Acute Respiratory Syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160:509–516, 2004.

[55] X. Wan and G. E. Karniadakis. An adaptive multi-element generalized polynomial chaos method for stochastic differential equations. *Journal of Computational Physics*, 209:617–642, 2005.

[56] G. R. Warnes. *mcgibbsit: Warnes and Raftery's MCGibbsit MCMC diagnostic*, 2005. R package version 1.0.5.

[57] N. Wiener. The homogeneous chaos. *American Journal of Mathematics*, 60:897–936, 1938.

[58] D. Wilkening. Sverdlovsk revisited : Modeling human inhalational anthrax. *Proceedings of the National Academy of Science*, 103:7589–7594, 2006.

[59] D. Xiu and G.E. Karniadakis. Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos. *Computer Methods in Applied Mechanics and Engineering*, 191:4927–4948, 2002.

[60] Y. Yang, M. E. Halloran, J. D. Sugimota, and I. M. Longini. Detecting human-to-human transmission of avian influenza A (H5N1). *Emerging Infectious Diseases*, 13(9):1348–1353, 2007.

This page intentionally left blank.

Sandia National Laboratories