

Introduction

Dexter Barrows

March 22, 2016

Epidemic forecasting is an important tool that can help inform public policy and decision-making in the face of an infectious disease outbreak. Successful intervention relies on accurate predictions of the number of cases, when they will occur, and where. Without this information it is difficult to efficiently allocate resources, a critical step in curbing the size and breadth of an epidemic.

Despite the importance of reliable forecasts, obtaining them remains a challenge from both a theoretical and practical standpoint. Mathematical models can capture the essential drivers in disease dynamics, and extend them past the present into the future. However, different epidemics may present with varying dynamics and require different model parameters to be accurately represented. These parameters can be inferred by using statistical model fitting techniques, but this can become computationally intensive, and the modeller risks “overfitting” by attempting to capture too many drivers with too little data. Thus, The modeller must exercise restraint in model selection and fitting technique.

Securing precise, error-free data in the midst of an outbreak can be difficult if not impossible, thus uncertainty in what we observe in building mathematical models of disease spread must be accounted for from the get-go. Models must differentiate between natural variation in the intensity of disease spread (process error) and error in data collection (observation error) in order to accurately determine the dynamics underlying a data set, adding another layer of complexity. With these caveats and concerns acknowledged, we can turn to a discussion of technique.

Broadly, there are three primary categories of techniques used in forecasting: phenomenological, pure mechanistic, and semi-mechanistic.

Phenomenological methods operate purely on data, fitting models that do not try to reconstruct disease dynamics, but rather focus purely on trend. A long-standing and widely-used example is the Autoregressive Integrated Moving Average (ARIMA) model. ARIMA assumes a linear underlying process and Gaussian error distributions. It uses three parameters representing the degree of autoregression (p), integration (trend removal) (d), and the moving average (q), where the orders of the autoregression and the moving average are determined through the use of an autocorrelation function (ACF) and partial autocorrelation function (PACF), respectively, applied to the the data *a priori*.

Pure mechanistic approaches simply try to capture the essential drivers in the disease spreading process and use the model alone to generate predictions. For example one could use a compartment model in which individuals are divided into categories based on whether they are susceptible to infection or infected but not yet themselves infectious, infectious, or re-

covered. These models are referred to as susceptible-infectious-removed (SIR) models and are heavily used in epidemiological study. Typically the transition between compartments is governed by a set of ordinary differential equations, such as

$$\begin{aligned}\frac{dS}{dt} &= -\beta IS \\ \frac{dI}{dt} &= \beta IS - \gamma I \\ \frac{dR}{dt} &= \gamma I,\end{aligned}\tag{1}$$

where S , I , and R are the number of individuals in each compartment, β is the “force” of infection acting on the susceptible population, and γ is a recovery rate. As an outbreak progresses, individuals transition from the susceptible compartment, through the infectious compartment, then finish in the removed compartment where they no longer impact the system dynamics. Many extensions of the SIR model exist and are commonly used, such as the SEIR model in which susceptible individuals pass through an exposed class (or several) where they have been infected but are not yet themselves infectious, and the SIRS model in which individuals become susceptible again after their immunity wanes.

Combining the phenomenological and mechanistic approaches are the semi-mechanistic techniques. These methods use a model to define the expected underlying dynamics of the system, but integrate data into the model in order to refine estimates of the model parameters and produce more accurate forecasts. Typically the first step in implementing such a technique is fitting the desired model to existing data. There are many ways to do this, most of which fall into two main categories: particle filter-based (PF) methods, and Markov chain Monte Carlo-based (MCMC) methods. From there data can either be integrated into the model by refitting the model to the new longer data set, or in an “on-line” fashion in which data points can be directly integrated without the need to refit the entire model. Normally, MCMC-based machinery must refit the entire model whereas PF-based approaches can sometimes integrate data in an on-line fashion.

Another, broader, distinction among techniques can be drawn between those that rely on assumptions of linearity, and those that make no such assumption. As epidemic dynamics are highly non-linear, it can be questionable as to even consider linear approaches to epidemic forecasting at all. In particular, stalwart approaches such as ARIMA and the venerable Kalman filter face a distinct (at least theoretical) disadvantage in the face of newer PF-based methods. Additionally, these methods are very-well-studied, and further work showing their viability would likely prove extraneous in the modern academic landscape.

Somewhat frustratingly, there exists no “gold standard” in forecasting. As methodology varies widely in theoretical justification, implementation, and operation, it is difficult to compare the state of the art in forecasting methods from a first-principles perspective. Further, published work making use of any of these methods to forecast uses different prediction

accuracy metrics, such as SSE, peak time/duration/intensity, correlation tests, or RMSE, among others. Thus it is difficult to select the best tool for the job when faced with a forecasting problem.

The primary focus of this work is to compare best-in-class methods for forecasting in several epidemically-focused scenarios. These include the a “standard” one-shot forecast outbreak in which the outbreak subsides and does not recur, a seasonal outbreak scenario such as the one we see with influenza each year, and a spatiotemporal scenario in which multiple spatial locations are connected and disease is free to spread from one to another.

From MCMC-based methods we have selected Hamiltonian MCMC [*Neal ref*], a (slightly) less cutting-edge but nonetheless highly effective technique. We are using HMCMC through an implementation in the R package RStan [*RStan ref*], which at its core uses HMCMC, but also contains implementations of several other innovative techniques. Interestingly, the original goal of this package was not to implement a statistical programming language similar to Just Another Gibbs Sampler (JAGS) [*JAGS ref*] or Bayesian Inference Using Gibbs Sampling (BUGS) [*BUGS ref*], but with an HMCMC backend. In fact the developers’ original goal was to implement any method that could fit multilevel hierarchical models without halting as they were witnessing with BUGS and JAGS. Only after experimenting with several options and starting to hear about it more and more frequently did they attempt to work with HMCMC. In the end, the scope of the project grew to include the development and subsequent integration of the No-U-Turn Sampler (NUTS), and an implementation of automatic differentiation machinery.

For PF-based methods we have selected IF2 [*Ionides ref*], a very recently developed approach that uses multiple particle filtering rounds to generate Maximum Likelihood Estimates (MLEs). It functions similarly to its predecessor, the Maximum likelihood via Iterated Filtering (MIF) algorithm, but aims to be simpler, faster, and more accurate. Theoretical justification and synthetic testing indicates that IF2 meets these goals, and as such the authors recommend skipping MIF and jumping straight to IF2 if an algorithm of their variety is sought. And so, we are doing just that. We wrote our own IF2 implementation in C++ and integrated it into R using the Rcpp package [*Rcpp ref*]. The developers of MIF and IF2 have their own R package that implements MIF and IF2, Partially Observed Markov Processes (POMP), but it didn’t provide some of the diagnostic information we needed, so it was not used here.

Finally, from the phenomenological methods we have selected the sequential locally weighted global linear maps (S-map) [*Sugihara ref*], combined with Dewdrop Regression [*Sugihara ref 2*]. These methods stand on their own as a unique take on the forecasting problem, and bear little resemblance to other methodology. The virtues of these techniques have been long extolled by their developers, but their efficacy when compared to competing methods has not been well-studied. This work will mark one of the first times this has been done.

This paper will begin with descriptions of HMCMC and IF2 with examples of simple model fittings in Chapters 2 and 3. Chapter 4 explores parameter fitting of a stochastic

SIR model to synthetic data. Chapter 5 will establish the full forecasting frameworks used with IF2 and HMCMC, and compare them in a simple scenario. All three methods will be used to compare forecasts using a SIRS model in Chapter 6. Chapter 7 will show forecasts using the aforementioned IF2 and HMCMC frameworks, along with Dewdrop Regression combined with S-mapping. Finally, a summary of these results, and a discussion of parallel computing and future directions will finish the paper in Chapter 8.