Models, Prediction, and Estimation of Outbreaks of Infectious Disease

Peter J. Costa
James P. Dunyak
Mojdeh Mohtashemi
{pjcosta@mitre.org, jdunyak@mitre.org, mojdeh@mitre.org}
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730–1420

Abstract

Conventional SEIR (Susceptible–Exposed–Infectious–Recovered) models have been utilized by numerous researchers to study and predict disease outbreak. By combining the predictive nature of such mathematical models along with the measured occurrences of disease, a more robust estimate of disease progression can be made. The Kalman filter is the method designed to incorporate model prediction and measurement correction. Consequently, we produce an SEIR model which governs the short term behaviour of an epidemic outbreak. The mathematical structure for an associated Kalman filter is developed and estimates of a simulated outbreak are provided

1. Introduction

Mathematical models have been used to study the outbreak of a number of infectious diseases [1, 2, 6]. In particular, difference and differential equations are the methodologies in which such models are written [4, 5, 6]. Many research hospitals and/or public health departments are maintaining a database of emergency room visits by patients with categorized complaints. The combination of a mathematical model of an outbreak with daily measurements beckons the application of a Kalman filter to provide an optimal estimate of the number of infections. This paper will provide the mathematical infrastructure required to implement a Kalman filter on simulated emergency room data.

The program of this discussion will be to provide a general model, discuss model simplification, and demonstrate the efficacy of the filter on simulated data. In this first section, we establish common notation and a general model for the outbreak of a specific (but unknown) infectious disease through a general population.

1.1 Notation

$S = S(t)$ = number of people in the population susceptible to the disease at time t
$E = E(t)$ = number of people in the population exposed/infected by the disease at time t
$I = I(t)$ = number of people in the population who are infectious at time t
$R = R(t)$ = number of people in the population who have recovered from the disease at time t
There are a number of parameters which will need to be either modeled or estimated from the data. It is assumed that these parameters are time invariant though more sophisticated efforts and information could produce time–varying models. A description of these parameters is listed below.

1.2 Parameters

β = probability of disease transmission
v = rate of seroconversion (i.e., from exposed to infectious)
μ_I = death rate of infectious due to the disease
α = recovery delay rate
$\rho(I) = \beta I(t)$ = conversion rate from susceptible to exposed/infected (also called the *force of infection*)

In figure 1 below, a schematic diagram expresses the graphical representation of the spread of an infectious disease through a population. Implicit in this figure is the assumption that everyone in the population is susceptible to the disease. The first boxes illustrate the migration of the population of susceptibles $S(t)$ to those exposed and infected $E(t)$. The rate at which the susceptibles are infected is proportional to the number of contacts c with the infectious population $I(t)$ times the probability of disease transmission per contact β times the proportion of the population which is

0-XXXX-XXXX-X/05/$20.00 ©2005 IEEE

infectious: $\rho(I) = \beta I(t)$. Since the infected leave the population of susceptibles a negative sign is attached to this quantity. Consequently, $dS(t)/dt = 0 - \rho(I)S(t) \equiv \beta I(t)S(t)$. In a similar manner, the disease dynamics of equation (1.1) are formed.

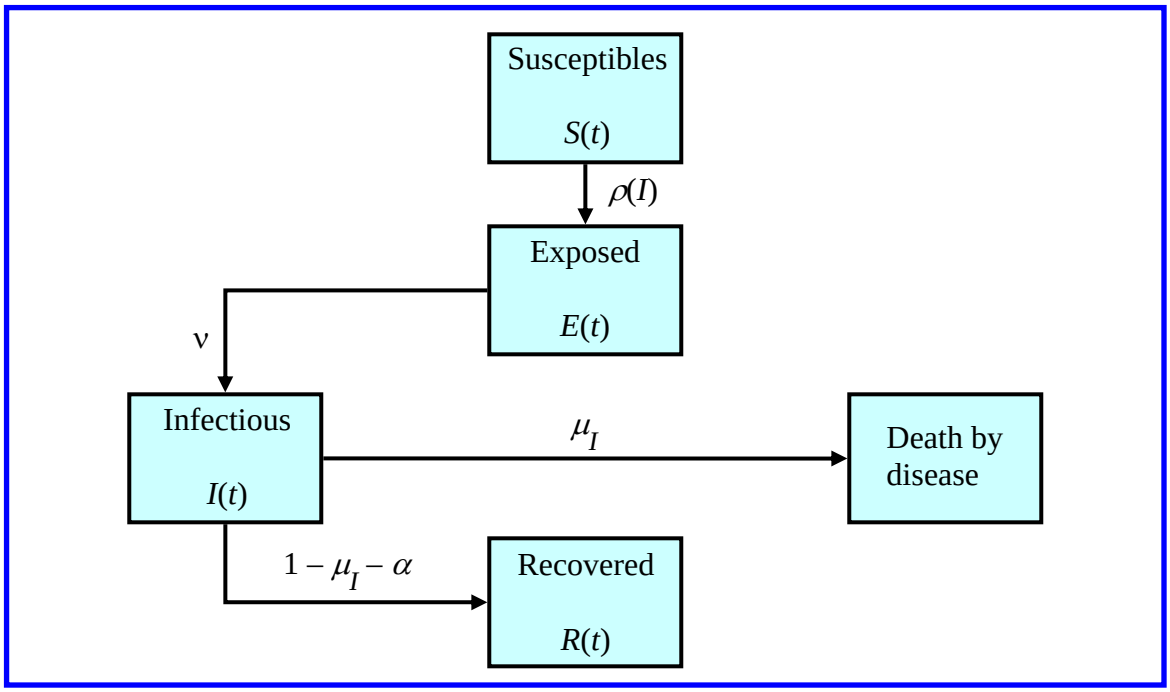

Figure 1.1. Disease dynamics

1.3 Disease Dynamics

$$\begin{aligned}
\frac{dS}{dt} &= -\rho(I)S(t) \equiv -\beta I(t)S(t) \\
\frac{dE}{dt} &= \rho(I)S(t) - \nu E(t) \equiv \beta I(t)S(t) - \nu E(t) \\
\frac{dI}{dt} &= \nu E(t) - \mu_I I(t) - (1-\mu_I - \alpha)I(t) \\
&\equiv \nu E(t) - (1-\alpha)I(t) \\
\frac{dR}{dt} &= (1-\mu_I - \alpha)I(t)
\end{aligned} \qquad (1.1)$$

This "full" model expressed in (1.1) operates under the simplifying assumption of a sufficiently short time–scale such that no significant population enters the susceptible population and that the parameters β, ν, μ_I, and α do not vary with respect to time. The efforts behind this work are to present a model for a short time–scale within the epidemic cycle (i.e., on the order of 2–3 weeks). Consequently, a series of simplifying assumptions can be made which are listed below.

Assumptions

(i) Short time–scale: $t \in [t_o, t_o + \Delta t]$ where the change in time Δt is less than three weeks.

(ii) No immigration to or emigration from the subpopulations

(iii) Insufficient time for R (recovereds) to return to the population of susceptibles

(iv) For $t \in [t_o, t_o + \Delta t]$, $S(t) = S(t_o) = S_o$.

From (iv), $\frac{dS}{dt} = 0$ and $S(t) = S_o$ (constant). Set $\rho(I) = \beta S_o I(t) \equiv \rho_o I(t)$, where $\rho_o \equiv \beta S_o$, so that the second and third equations of the disease dynamics become

$$\begin{aligned}
\frac{dE}{dt} &= \rho_o I(t) - \nu E(t) \\
\frac{dI}{dt} &= \nu E(t) - (1-\alpha)I(t)
\end{aligned} \qquad (1.2)$$

Observe that the fourth equation of the disease dynamics is completely decoupled from the middle two equations. Consequently, the population of recovereds can be computed as

$$R(t) = R(t_o) + (1-\mu_I-\alpha)\int_{t_o}^{t} I(\tau)d\tau\,. \qquad (1.3)$$

By setting $\mathbf{X} = [E,I]^T$, the reduced set of disease dynamics can be written in the vector–matrix form

$$\frac{d\mathbf{X}}{dt} = A\mathbf{X} \qquad (1.4)$$

where $A = \begin{bmatrix} -\nu & \rho_o \\ \nu & 1-\alpha \end{bmatrix}$.

The measurements of this system are a portion of the number of infectious which report to emergency rooms on a day–to–day basis. More precisely, let T be the probability that a member of the infectious population appears in a reporting emergency room. Then, the measurements are

$$m(t) = TI(t). \qquad (1.5)$$

The measured quantity, $TI(t)$, rather than the modeled population of infectious people $I(t)$, is what emergency departments reported. Thus, make the following change of variables (1.6) to transform the problem to a "non–dimensional" framework.

$$\begin{aligned} I(t) &\mapsto TI(t) \equiv \hat{I}(t) \\ E(t) &\mapsto TE(t) \equiv \hat{E}(t) \end{aligned} \qquad (1.6)$$

Since, $\frac{d\hat{E}}{dt} = T\frac{dE}{dt}$ and $\frac{d\hat{I}}{dt} = T\frac{dI}{dt}$, then multiplying (1.2) by T and simplifying yields the "dimensionless" disease dynamics

$$\begin{aligned} \frac{d\hat{E}}{dt} &= \rho_o\hat{I}(t) - \nu\hat{E}(t) \\ \frac{d\hat{I}}{dt} &= \nu\hat{E}(t) - (1-\alpha)\hat{I}(t) \end{aligned} \qquad (1.7)$$

and the associated measurements

$$m(t) = \hat{I}(t)\,. \qquad (1.8)$$

Now with $\hat{\mathbf{X}} = \left[\hat{E},\hat{I}\right]^T$ and A as above, the disease dynamics can be written as

$$\frac{d\hat{\mathbf{X}}}{dt} = A\hat{\mathbf{X}} \qquad (1.9)$$

where $\hat{X}$ is the *state vector*. As equation (1.9) illustrations, the disease dynamics are linear. Moreover, there are regular time measurements (1.8). Modern control theory was developed around this very scenario: The need to solve linear differential equations in association with regularly sampled (in time) measurements. An optimal estimate of the model predicted/ measurement corrected state of a disease outbreak can be obtained via the Kalman filter. The discussion is hereafter, framed in the Kalman filter context.

2. The Kalman Filter

Since the mathematical models of the disease dynamics (1.9) and measurements (1.8) are inherently imperfect, "noise" in the form of zero–mean Gaussian random processes are added to enhance these modelling deficiencies. Thus, to the state dynamics, add a vector $\mathbf{w}(t) \sim N(\mathbf{0},Q(t))$ called the state or *system error*. The matrix $Q(t)$ is called the state or *system noise covariance*. Similarly, to compensate for the variability in the measurements, a vector $\mathbf{v}(t) \sim N(\mathbf{0},V(t))$ called the *measurement error* is added (1.8). The matrix $V(t)$ is called the *measurement noise covariance*. The definitions below help to develop the Kalman filter (see, e.g., Costa [3]).

State Vector: $\hat{\mathbf{X}} = \begin{bmatrix} \hat{E} \\ \hat{I} \end{bmatrix}$

State Dynamics: $\frac{d\hat{\mathbf{X}}}{dt} = A\hat{\mathbf{X}}$

System Model: $\frac{d\hat{\mathbf{X}}(t)}{dt} = A\hat{\mathbf{X}}(t) + \mathbf{w}(t)$

System Noise Covariance: $Cov\left[\mathbf{w}(t)\right] \equiv Q(t)$

Measurement: $m(t) = \hat{I}(t) \equiv H\hat{\mathbf{X}}(t)$

Measurement model: $m(t) = H\hat{\mathbf{X}}(t) + \mathbf{v}(t)$,

Measurement Jacobian: $H = [0,1]$

Measurement Noise Covariance:

$$Cov\left[\mathbf{v}(t)\right] \equiv V(t)$$

Transition matrix: $\Phi(t,t_o) = \exp\left(A[t-t_o]\right)$

where $A = \begin{bmatrix} -\nu & \rho_o \\ \nu & 1-\alpha \end{bmatrix}$

Measurement Ensemble to time t_n:

$$M_n = \{m(t_1), m(t_2), \cdots, m(t_n)\}$$

$$M_o = \{\} = \text{the empty set}$$

State Prediction:

$$\vec{\mathbf{X}}_p(t_k, M_{k-1}) = \Phi(t_k, t_{k-1})\vec{\mathbf{X}}_p(t_{k-1}, M_{k-1})$$

$$\vec{\mathbf{X}}_p(t_1, M_o) = \Phi(t_1, t_o)\vec{\mathbf{X}}_p(t_o, M_o) \equiv \Phi(t_1, t_o)\vec{\mathbf{X}}(t_o)$$

State Jacobian: $F = A$

Covariance dynamics (Ricatti Equation):

$$\frac{dP(t)}{dt} = P(t)F^T + FP(t) + Q(t)$$

Covariance prediction:

$$P(t_n) = \Phi(t_n, t_{n-1})P(t_{n-1})\Phi^T(t_n, t_{n-1}) + \int_{t_{n-1}}^{t_n} \Phi(t_n, s)Q(s)\Phi^T(t_n, s)ds$$

Kalman gains matrix:

$$K(t_n) = P(t_n)H^T\Im(t_n)$$

Information matrix:

$$\Im(t_n) = [HP(t_n)H^T + V(t_n)]^{-1}$$

State correction:

$$\vec{\mathbf{X}}_c(t_n, M_n) = K(t_n)[m(t_n) - m_p(t_n)]$$

Predicted measurement:

$$m_p(t_n) = \hat{I}_p(t_n, M_{n-1})$$

State estimate:

$$\hat{\mathbf{X}}(t_n, M_n) = \vec{\mathbf{X}}_p(t_n, M_{n-1}) + \vec{\mathbf{X}}_c(t_n, M_n)$$

Covariance update (Joseph form):

$$P(t_n) = [I_{nxn} - K(t_n)H]P(t_n)[I_{nxn} - K(t_n)H]^T + K(t_n)V(t_n)K^T(t_n)$$

3. Simulation

A mathematical model that simulates the underlying dynamics of the hospital daily visits that are influenza related was developed in the form of equation (3.1)

$$D(t) = 2\cos(2\pi t/365) + 8 + w_t\,. \qquad (3.1)$$

Here t = 0, 1, 2, … , 5x365 is measured in single days over five years, and $w_t \sim N(0,2)$ is normally distributed noise. We assumed the following set of initial conditions and parameters:

$S_o = 1000, \vec{E}_o = 10, \vec{I}_o = 1, R_o = 2,$

$\nu = 0.4, \beta = 0.5, \alpha = 0.3,$ and $\mu_I = 0.1$.

The system noise covariance Q was selected as a 10% variation of the initial state covariance $P(t_o) = (\vec{X}(t_o) - \vec{\mu}_o)(\vec{X}(t_o) - \vec{\mu}_o)^T$ and $\vec{\mu}_o = \frac{1}{2}\left(\vec{E}_o + \vec{I}_o\right)$. Finally, the measurement noise covariance V was selected as the variance in the data.

The filter was run over the simulated data (3.1) to establish a baseline estimate of the number of exposed/infected and infectious reporting to an emergency department. The results are depicted in Figure 3.1 below. A one–standard deviation neighborhood, based on the estimated covariance matrices $P(t)$ was computed for the infectious class; see top portion of Figure 3.2. Then a simulated one–week (i.e., seven day) outbreak was introduced into the population at a random seed time t_o in the form of (3.2).

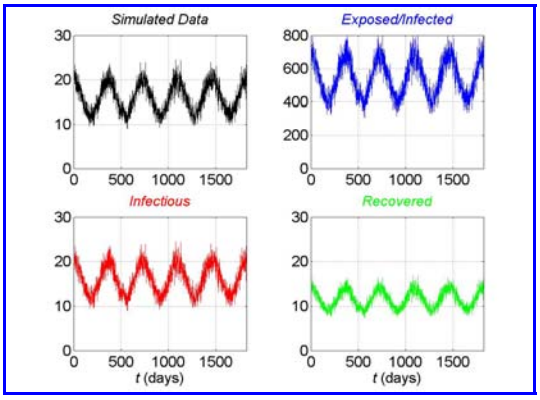

Figure 3.1. Baseline Kalman filter estimates from simulated data

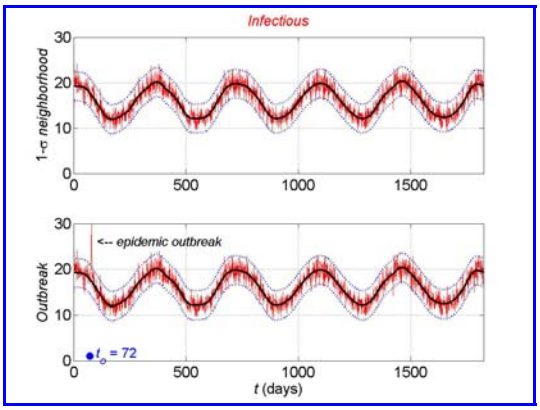

Figure 3.2. One standard deviation neighborhood of the infectious class

$$f_{outbreak}(t,t_o)=\begin{cases}0 \text{ for } t<t_o\\ 2(t-t_o) \text{ for } t_o\le t\le t_o+6\\ 0 \text{ for } t>t_o+6\end{cases} \quad (3.2)$$

That is, $f_{outbreak}(t,t_o)$ was added to the simulated data $D(t)$ in (3.1). If the Kalman filter estimate of the infectious class $\hat{I}(t_k,M_k)$, reflecting the influence of the measurements $D(t)+f_{outbreak}(t,t_o)$ through time t_k, exceeded the one–standard deviation neighborhood established for the baseline case within ten days of the start of the outbreak (i.e., for $t_k\in[t_o,t_o+10]$), then a *true positive* for outbreak detection was recorded. Otherwise, a *false negative* was recorded. To insure a sufficient number of measurements were processed by the Kalman filter, the range of the random outbreak time was restricted: $t_o\in[50,1800]$ days. One thousand random outbreaks were tested and the number of true positives (T_p) and false negatives (F_n) were recorded. For this test, 100% of the outbreaks were discovered within the requisite time period (10 days). In particular, 2.9% of the outbreaks were detected on day 2, 17.3% were detected on day 3, 59.8% were detected on day 4, 19.9% were detected on day 5, and 0.1% were detected on day 6 of the outbreak.

Summary

A set of mathematical models governing the outbreak of an infectious disease have been detailed. Simulated data have been generated. The associated Kalman filter has been developed and tested against the simulated data with positive results. Analysis concerning the variation of the model parameters and their effect upon the Kalman filter estimates and the application of this method to real recorded emergency department data will be the focus of future work

References

[1] R. M. Anderson and R. M., May, *Infectious Diseases of Humans*, Oxford Science Publications, 1992

[2] N. G. Becker and L. R. Egerton, "A Transmission Model of HIV", Mathematical Biosciences, Volume 119, 1994, pp. 205–224

[3] P. J. Costa, *Bridging Mind and Model*, St. Thomas Technology Press, St. Paul, MN, 1994

[4] G. Fulford, P. Forrester, and A. Jones, *Modelling with Differential and Difference Equations*, Cambridge University Press, 1997

[5] S. Gupta, R. M. Anderson, and R. M.,May, "Mathematical Models and the Design of Public Health Policy: HIV and Antiviral Therapy", SIAM Review, Volume 35, Number 1, March 1993, pp. 1–16

[6] J. D. Murray, *Mathematical Biology, Second, Corrected Edition*, Springer–Verlag, Berlin, 1993