Deriving a model for influenza epidemics from historical data

S. Lefantzi and J. Ray

![Sandia National Laboratories]

Deriving a model for influenza epidemics from historical data

S. Lefantzi and J. Ray

Sandia National Laboratories,

P. O. Box 969, Livermore CA 94551

{slefant,jairay@sandia.gov}

Abstract

In this report we describe how we create a model for influenza epidemics from histori-
cal data collected from both civilian and military societies. We derive the model when
the population of the society is unknown but the size of the epidemic is known. Our
interest lies in estimating a time-dependent infection rate to within a multiplicative
constant. The model form fitted is chosen for its similarity to published models for
HIV and plague, enabling application of Bayesian techniques to discriminate among
infectious agents during an emerging epidemic.

Acknowledgment

Contents

Appendix

Figures

Tables

This page intentionally left blank

1 Influenza Model

We assume that the probability of being infected (per unit time) changes in time, starting from a small value at the beginning of the epidemic and peaking when the medical counter-measures are put in place. We model the unknown infection rate (defined as the probability of being infected over a 24 hour period) as $q(s; k, r)$, where s is the time since the start of the epidemic (first infection) and (k, r) are parameters of the infection rate model

$$q(s; k, r) = \frac{r^k}{\Gamma(k)} s^{k-1} e^{-rs} \tag{1}$$

where $\Gamma(k)$ is a Gamma function. As can be seen, the expression $q(s; k, r)$ in Eq. 1 resembles a Gamma distribution. It is described only over the duration of the epidemic.

The infection rate, convolved with the cumulative distribution function of the incubation period of influenza provides us with a model for the evolution of the epidemic (see derivation in [3]):

$$\frac{n(T)}{N_{tot}} = \int_{s=0}^{T-\tau} q(s; k, r) F_I(T - s - \tau) ds \tag{2}$$

where:

$n(T)$ is the number of people who have exhibited symptoms by time T

τ is the time when the epidemic is believed to have started, and

N_{tot} is the total number of victims of the epidemic over its entire course.

For F_I we use the following form:

$$f_I(Y) = \frac{1}{2} \text{erfc} \left(-\frac{\ln(Y) - \mu}{\sigma \sqrt{2}} \right) \tag{3}$$

which denotes a log-normal distribution for the incubation period Y. Here μ and σ are derived from the mean and standard deviation of the influenza incubation period, taken to be 2 and 1 day respectively (as in Bombardt and Brown [1]).

Data from an epidemic is available as the number of sick people observed over a set observation period (sometimes a 24-hour period, but generally weekly) and the time-series is available over the entire duration of the epidemic, which is generally more than a month.

1.1 Fitting the Model to Data

We collected data from [1], [2] and [4]. They represent the number of sick people observed on a daily basis in [1] and on a weekly basis in [2] and [4]. For each of the

datasets in these papers, we extract a time-series of people reporting sick or exhibiting symptoms during the outbreak. The data is usually collected on a daily or weekly basis. In terms of Eq. 2, the time-series corresponds to $(n(T_i) - n(T_i - \delta))$, where T_i are the end-points of each interval in the time-series and δ is the width of the interval over which the data is collated (usually a day or a week). The model is fitted to data by changing k, r and τ, inside a gradient-based optimization loop that minimizes the mean-square difference between the model and (daily/weekly-recorded) observation time-series. The code was written in *MATLAB* and *fminsearch* was the optimization function used. The code for fitting a model to data collected from military populations can be found in Appendix A.

1.2 Influenza Results

In Fig. 1 we illustrate our parameter model fitting results for military data obtained from the 1918 flu epidemic at Camp Custer [1]. The circles represent the data, *i.e.* daily arrival of sick individuals, the solid line represents our influenza model fitted to the data (Eq. 2) and the dashed line (axis on the right) our infection rate model (Eq. 1).

In Fig. 2 we illustrate our parameter model fitting results for civilian weekly data obtained from CDC for $2004 - 05$ [2]. Here the circles represent the data, *i.e.* weekly sick arrivals, the solid line represents our flu model fitted to the data (Eq. 2) and the dashed line (axis on the right) our infection rate model (Eq. 1).

In Table 1 we present a summary of our results for parameters k, r, and τ fitted for each of the considered historic data sets. We also report the data type (military or civilian) and the daily or weekly time rate for sick arrivals.

Table 1: Model fitted parameters k, r, and τ for each historic data set.

TYPE	SOURCE	δ	k	r	τ
Military	Camp Custer, 1918 [1]	Daily	24.69	1.11	-0.19
Military	3rd Regiment,1918 [1]	Daily	17.33	1.39	0.03
Military	78th Regiment, 1918 [1]	Daily	22.60	1.54	-0.27
Military	6th Brigade, 1918 [1]	Daily	30.32	1.14	-6.87
Civilian	2001-02 US Influenza season, [2]	Weekly	25.37	0.18	0
Civilian	2002-03 US Influenza season, [2]	Weekly	22.13	0.16	0
Civilian	2004-05 US Influenza season, [2]	Weekly	23.03	0.17	0
Civilian	2006-07 US Influenza season, [2]	Weekly	22.14	0.15	-1.63
Civilian	2007-08 US Influenza season, [2]	Weekly	39.67	0.27	-4.00
Civilian	Catalunya, 2008 [4]	Weekly	40.49	0.35	0
Civilian	Catalunya, 2009 [4]	Weekly	22.25	0.17	0

Fig. 3 is a scatter plot of our results for k, r. Each symbol is a (k, r) pair obtained by fitting our model to a specific historic data set. The squares correspond to military data fits and the

diamonds correspond to civilian data fits. In our infection rate model (see Eq. 1) parameter *k* essentially describes how quickly the infection rate varies in the beginning of the epidemic (*i.e.* when it is increasing); our results do not show a clear differentiation between military and civilian societies in their response. Parameter *r* determines how quickly the infection rate decays as the epidemic is brought under control. We can see that the military data clusters around larger values of *r* with $(1.1 < r < 1.6)$ while the civilian data results in a clustering at the lower end $(0.1 < r < 0.5)$. Based on this clear clustering in *r*, it seems that military societies bring flu under control quicker.

In Fig. 4 we present the histograms for parameters *k* and *r* respectively, as listed in Table 1. We see no clear trend in *k* (Fig. 4 on the left), but the bulk of the values lie between 22 and 26 (mean $\mu_k = 26.4$, standard deviation $\sigma_k = 7.4$, median $= 23$); on the other hand *r* shows a clear multimodal distribution (Fig. 4 on the right), reflecting the difference in the response of civilian and military societies that was seen in Fig. 3.

In Table 2 we tabulate *consensus* values, appropriate for modeling influenza in Bayesian inference problems. These may serve as prior beliefs. Note that *k* has some outliers ($k \approx 40$) and consequently, the median provides a better representation. Note, too, that the coefficient of variation (σ_k/μ_k) for k is around 0.3. Methods for discriminating causes of emerging epidemics, based on these parameters are presented in [5].

Table 2: Summary of the influenza model. $N(\mu, \sigma)$ refers to a normal distribution.

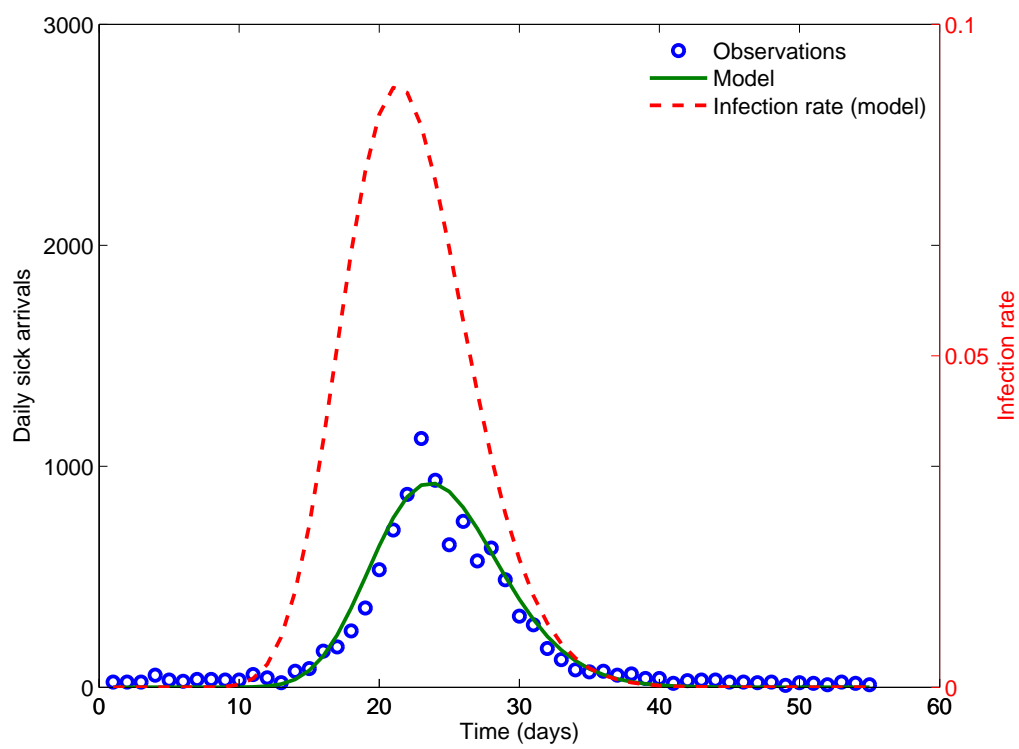| Epidemic model	Eq. 2
Incubation period	Log-normal; $\mu = 2$ days, $\sigma = 1$ day; Eq. 3
Infection rate model	Eq. 1
k	23
r	$N(1.3, 0.2)$ [military]; $N(0.21, 0.07)$ [civilian]

Figure 1: Model fitting to Camp Custer data from the 1918 influenza epidemic [1]
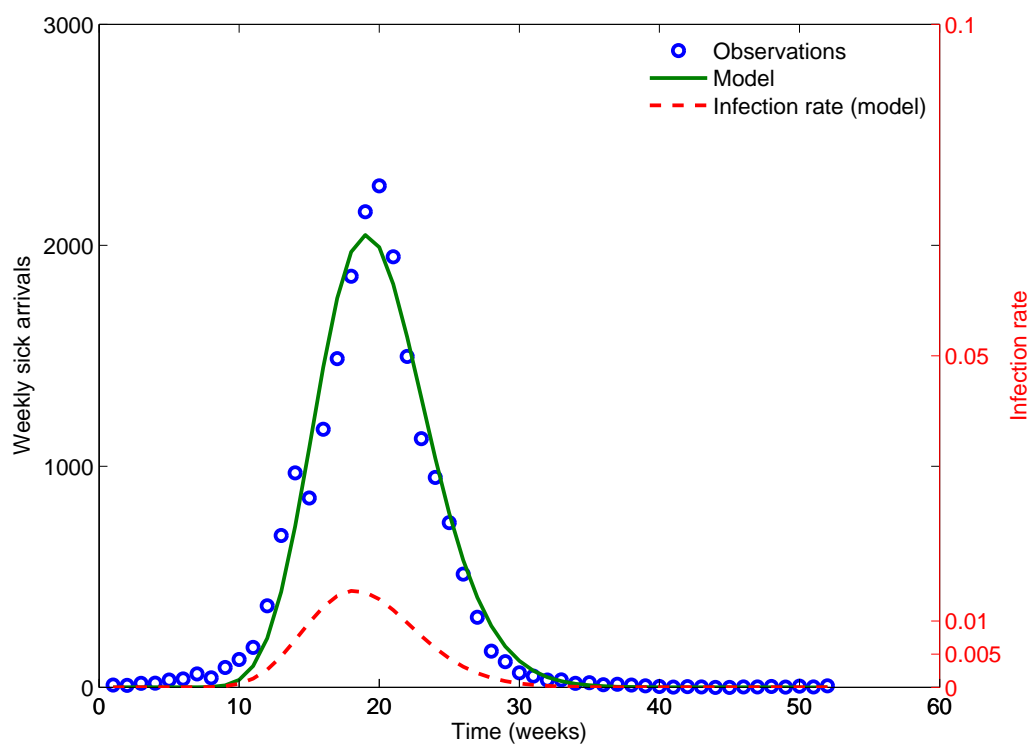
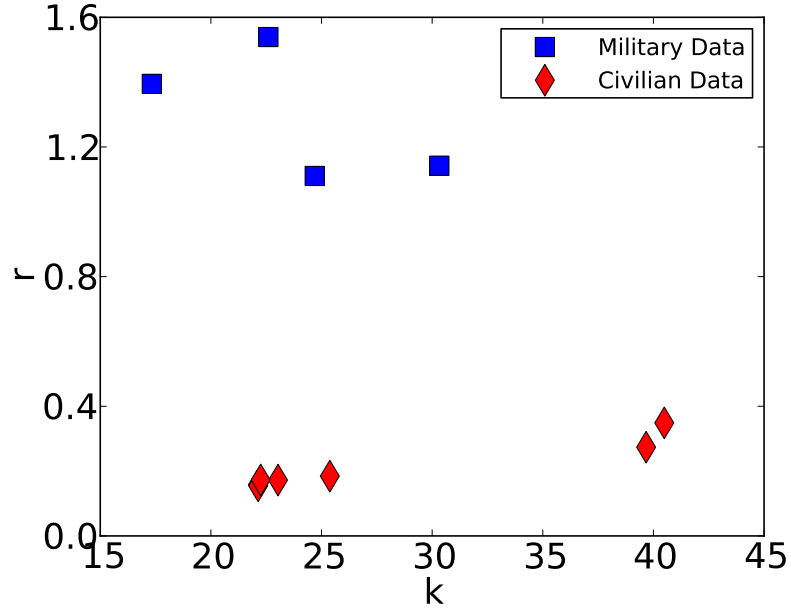Figure 2: Model fitting to CDC influenza data for 2004-05 [2]

Figure 3: (k, r) pairs obtained with our model fitting process for each data set
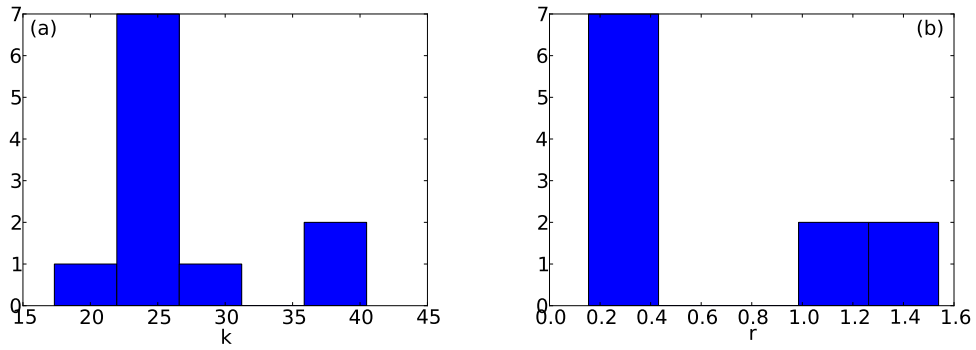

Figure 4: Histograms of parameters k (left) and r (right)

2   Conclusions

We have developed models for the progression of influenza in human populations. The model is framed as a integral, and predicts the number of people who exhibit symptoms and seek care over a given time-period. The start and end of the time period form the limits of integration. The disease progression model, in turn, contains parameterized models for the incubation period and a time-dependent infection rate. The incubation period model is obtained from literature, and the parameters of the infection rate are fitted from historical data including both military and civilian populations.

The calibrated infection rate models display a marked difference in which the 1918 Spanish Influenza pandemic differed from the influenza seasons in the US between 2001–2008 and the progression of H1N1 in Catalunya, Spain. The data for the 1918 pandemic was obtained from military populations, while the rest are country-wide or province-wide data from the twenty-first century. We see that the initial growth of infection in all cases were about the same; however, military populations were able to control the epidemic much faster i.e., the decay of the infection-rate curve is much higher. It is not clear whether this was because of the much higher level of organization present in a military society or the seriousness with which the 1918 pandemic was addressed. Each outbreak to which the influenza model was fitted yields a separate set of parameter values.

We suggest "consensus" parameter values for military and civilian populations in the form of normal distributions so that they may be further used in other applications. Representing the parameter values as distributions, instead of point values, allows us to capture the uncertainty and scatter in the parameters. Quantifying the uncertainty allows us to use these models further in inverse problems, predictions under uncertainty and various other studies involving risk.

This page intentionally left blank

References

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

[2] Centers for Disease Control and Prevention, Seasonal Influenza (Flu), Past Weekly Surveillance Reports. http://www.cdc.gov/flu/weekly/pastreports.htm; accessed in February 2011.

[3] 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.

[4] 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.

[5] Cosmin Safta, Jaideep Ray, Khachik Sargsyan, Sophia Lefantzi, Karen Cheng, and David Crary. Real-time characterization of partially observed epidemics using surrogate models. *Mathematical Biosciences*. Paper submitted September 2011.

This page intentionally left blank

A Matlab code for fitting an Influenza model to data collected from military populations

A.1 flu_driver.m

```
% flu_driver.m
% A script to test the epidemic subroutine
% Sophia Lefantzi, Sandia National Laboratories, 01/18/2011

clear ;    % Forget all variables
close all ; % Close all plot windows

% ---- User input
meanIncub = 2.0 ;
stdIncub  = 1.0 ;
% The input data file is a two column file. The fisrst column has the
% day number and the second column has the number of symptomatic/diagnosed
% people over that 24 hour period.
filename = 'CampCuster1918.dat' ;
plotIt = false ;

% Initial values for the parameters to be fitted to data
k = 5 ;
r = 1.0e-01 ;
tau  = -1 ;

% ---- End user-inputs

% get the flu params
varIncub = stdIncub^2 ;
[mu, sigma] = fluParams(meanIncub, varIncub) ;

% --- Read in data and find particulars about this dataset
D = load (filename) ;
ndays = size(D,1) ;
observations = D(:,2);
Ninf = sum(observations) ;


% ---- Run the model and see what the error wrt obs is
% Struct to contain all user parameters
UserData.mu = mu ;
UserData.sigma = sigma ;
```

```
UserData.Ndays = ndays ;
UserData.Ninf = Ninf ;
UserData.observations = observations;
UserData.PlotIt = plotIt ;

P0 = [k, r, tau] ;

% --- Do an optimization on P0 to get good values of P
options = optimset('TolX', 1.0e-1, 'MaxIter', 1000) ;
Pfinal = fminsearch(@(P) EpidemicError(P, UserData), P0, options) ;

% ---- Compute infection rate
s = 1:ndays ;
k = Pfinal(1) ;
r = Pfinal(2) ;
q = ( s.^(k-1) .* exp(-r.*s) * r^k ) / gamma(k);
tau = Pfinal(3) ;

% -- compute the model predictions
evolution=epidemic(k, r, mu, sigma, ndays, Ninf, tau) ;
evolution(2:ndays) = evolution(2:ndays) - evolution(1:ndays-1) ;
temp = [observations, evolution] ;
```

A.2 EpidemicError.m

```
function [ErrSqrt] = EpidemicError(P, UserData)
%EpidemicError Function to calculate the discrepancy between observations
%and results from model with guessed parameters
%   This function accepts some guessed values for the epidemic parameters
%   in P, namely (k, r, tau), runs the epidemic model and calculates the
%   sum-of-discrepancy-squared for P.
% ------------------------------------------------------------

% Sophia Lefantzi, Sandia National Laboratories, 01/18/2011

% ----Extract parameters from P.
k = P(1) ; r = P(2) ;  tau = P(3) ;

% ---- Extract user's dataset-specific info
mu = UserData.mu ;
sigma = UserData.sigma ;
ndays = UserData.Ndays ;
Ninf = UserData.Ninf ;
```

```
% Run the model. Results are in evolution.
%fprintf('P = [%f, %f, %f]\n', k, r, tau) ;
evolution =  epidemic(k, r, mu, sigma, ndays, Ninf, tau);
temp = evolution ;
for iday = 1 : (ndays-1)
    temp(iday+1) =  evolution(iday+1) - evolution(iday) ;
end
evolution = temp ;


 % ---- Compute the Squared Error
 EPS = 1.0e-6 ;
 Err = UserData.observations - evolution ;
 ErrNorm = norm(Err, 2) ;
 ErrSqrt = ErrNorm;
 fprintf('P = [%f, %f, %f], E = %f\n', k, r, tau, ErrSqrt) ;
 return;
end
```

A.3   epidemic.m

```
function [evolution] = epidemic(k, r, mu, sigma, ndays, Ninf, tau)
% epidemic(k,r,mu,sigma, ndays, Ninf) provides the evolution of the flu
% epidemic given k,r,mu and sigma.
%  This file returns an evolution of a flu epidemic, given (k, r) and (mu,
%  sigma).
%  Output:
%     evolution, a vector ndays long, with the epidemic evolution
%
%  Input :
%     - ndays is the number of days the epidemic evolves, - Ninf, total
%     size of the epidemic i.e., number of people who did turn sick over
%     the whole epidemic. - tau, number of days before first data, when the
%     epidemic started (first infection). tau < 0.
%
% -------------------------------------------------------------------------

% Sophia Lefantzi, Sandia National Laboratories, 01/18/2011
% ---- Allocate the array where we will store the evolution
evolution = zeros(ndays, 1) ;

% ---- Ensure that the parameters are inside desired bounds
```

19

```matlab
parametersAreGood = true ;
if ( (tau>0) || (k<0) || (r<0) )
    parametersAreGood = false ;
end

if (parametersAreGood == false)
    return;
end

% ---- Loop over all days and fill in evolution
for iday = 1 : ndays
    fluInt = quad(@(s)fluIntegral(s,k,r,mu,sigma,tau,iday),0,iday-tau) ;
    evolution(iday) = Ninf * fluInt ;
end

return ;
end


function [result] = fluIntegral(s,k,r,mu,sigma,tau,iday)

incubation  = iday-s-tau ;  % Quad sends in a vector of s

% a vector as long as incubation
cdfOfIncub = zeros(size(incubation)) ;

% --- make sure incubations are > 0
goodIndices = find( incubation > 0) ;
tmp = ( log(incubation(goodIndices))-mu ) / ( sigma*sqrt(2) ) ;
cdfOfIncub(goodIndices) = 0.5*erfc( - tmp ) ;
q = ( s.^(k-1) .* exp(-r.*s) * r^k ) / gamma(k);
result = cdfOfIncub .* q ;

return;
end
```

A.4 fluParams.m

```matlab
function [mu,sigma] = fluParams(mean_flu, variance_flu)
% fluParams Flu parameters given the mean and variance
% of incubation period
% Sophia Lefantzi, Sandia National Laboratories, 01/18/2011
% -----------------------------------------------
```

```
sigma = sqrt ( log(1+variance_flu/mean_flu^2) );
mu = log(mean_flu) - (sigma^2)/2 ;
return;
end
```

DISTRIBUTION:

1	Sophia Lefantzi, 08954	MS 9152
1	Jaideep Ray, 08954	MS 9159
1	Technical Library, 08944 (electronic)	MS 0899

Sandia National Laboratories