Ensemble Kalman Filter for Parameter Estimation of Epidemic Models Using Google Flu Trends

Peter Shaffery

December 9, 2016

1 Introduction

Epidemic surveillance and mathematical models of epidemic behavior are important tools for public health policy, and both are used by officials in preparing and implementing strategies to mitigate epidemic severity. In the United States flu epidemic surveillance is handled by the Center for Disease Control (CDC), who record incidence of "inluenza like illnesses" (ILI) at around 2400 sites in the US. The CDC aggregates and publishes the ILI incidence for 10 regions which comprise the US. While this reporting is very accurate, it can have a substantial delay of up to 2 weeks due to the preparation time and correction of erroneous reports. Flu seasons typically last around 36 weeks and the main period of flu spread can be substantially shorter, so a 2 week lag presents a severe obstacle for public health officials attempting to combat the epidemic.

Google Flu Trends was a service produced by Google working the CDC that ran from 2008-2013 that attempted to address this issue. With the rise in internet usage across the United States it was hoped that the frequency of certain search phrases in the popular

search engine Google were related to flu incidence in the US. A number of search terms were identified that were predictive of the ILI incidence in the US, and Google developed a logistic model for estimating the number of ILI cases in the US in a single week. These estimates were typically quite accurate, and were able to be published at the end of each week, 1-2 weeks ahead of the CDC reports. Unfortunately, in the 2012-2013 flu season Flu Trends overpredicted the ILI substantially and the project has since been (publicly) abandoned. While the Flu Trends project was not ultimately successful, it provides an important first step in improving public health surveillance. Treating it as a case study, future projects can hopefully identify the issues that overcame Flu Trends and account for them.

In the meantime it is useful to build tools that can utilize Flu Trends-type data to inform public health policy. One important challenge facing epidemic surveillance is using real-time data to predict epidemic severity; given the data up to some time t, need to decide if the epidemic is seasonal or a pandemic. Data assimilation is a natural choice to address this challenge, especially filters which provide parameter estimates, as well state estimates. One assimilation approach investigated by Dukic, Lopes, and Polson (2012) is the particle filter. The particle filter has some advantages that might be desirable for a flu surveillance network, however it struggles in higher dimensions so it may not be appropriate for flu surveillance of the 10 CDC regions individually.

In this project my goal was to construct a filter that accounted for each of the 10 regions and their interactions, that also provided parameter estimates. To this end I implemented an ensemble Kalman Filter, specifically the square root filter. This algorithm generally deals well with higher dimensional models, so I hoped that it would provide good inference for this larger model.

2 Epidemic Model

A simple, discrete time, multi-region SEIR epidemic model is given by:

$$\begin{aligned} s_{t+1} &= s_t - \beta s_t i_t / n + F s_t \\ e_{t+1} &= (1-\alpha) e_t + \beta s_t i_t / n + F e_t \\ i_{t+1} &= (1-\gamma) i_t + \alpha e_t \\ r_{t+1} &= r_t + \gamma i_t \end{aligned} \tag{1}$$

A notational note: all lower case letters indicate vectors, and any operation between vectors is elementwise. Upper case letters indicate matrices (with normal, matrix multiplication) and greek letters indicate parameters. At time t, s_t gives the number of susceptibles, e_t gives the number of latent individuals (sick, but not symptomatic or contagious), i_t is the number of infected individuals (symptomatic and contagious), and r_t is the recovered/removed category (either dead or recovered with immunity). The matrix F represents the inter-region flow rate. Each state vector in (1) has one element for each region

The matrix F was determined by a so-called gravity model, where $F_{ij} = \frac{m_i m_j}{d_{ij}}$. The parameter m_i are the 'mass' of the i^{th} region and the parameter d_{ij} is some measure of distance between regions i and j. In addition to tracking $s_t, e_t, i_t, r_t, \alpha, \beta,$ and γ I also track the m_i, but assume that d_{ij} is fixed and known.

Since the ILI is an index of flu incidence, and not an actual measure of it, I used its growth rate in a single week as an observation, and not the estimated ILI itself. From (1) we have that the evolution of the growth rate is given by $g_t = (i_{t+1} - I_t)/I_t$. Thus

the evolution and observation (y_t) equations of the system are:

$$\begin{aligned} y_{t+1} &= g_{t+1} + \epsilon^y_{t+1} \\ g_{t+1} &= -\gamma + \alpha\frac{e_t}{i_t} + \epsilon^g_{t+1} \\ i_{t+1} &= (1 + g_{t+1})i_t \\ s_{t+1} &= s_t - \beta s_t i_t/n + F s_t \\ e_{t+1} &= (1-\alpha)e_t + \beta s_t i_t/n + F e_t \\ r_{t+1} &= r_t + \gamma i_t \end{aligned} \tag{2}$$

Where $\epsilon^y_t \sim N(0, \sigma^2_y)$ and $\epsilon^g_t \sim N(0, \sigma^2_g)$.

3 Configuration

I ran the SRF with 100 ensemble members for the 36 weeks of data spanning the '03-'04 US flu season. The intial state variables were chosen by drawing $usim\text{Unif}(.99, 1)$ and then setting $s_0 = nu$, where n is the vector of population sizes in each of the 10 regions. The remaining initial states were set as $e_0 = r_0 = 0$ and $i_0 = n - s_0$. The growth rate was initially set to 0 and the initial parameter values were drawn from the following priors given in Table 1.

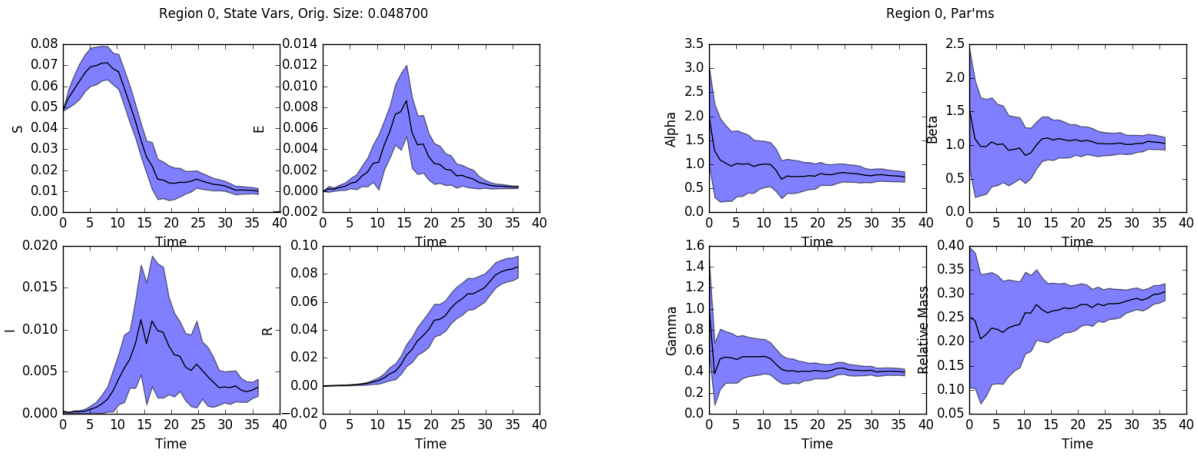
I had initially used some inflation for the perturbation matrix, but quickly abandoned it as it caused the inference to blow up.

4 Results

Here I present results for HHS Region 1, which happens to be New England. Fig. 1 shows filter output for the '03-'04 flu season. To verify that the filter was performing correctly I also tested it on a synthetic time series of growth rates with known parameters (and

Variable	Prior
α	$N(2,1)$
β	$N(1.5,1)$
γ	$N(1,.5)$
m_i	Unif$(0,3)$

Table 1: Parameter prior distributions.

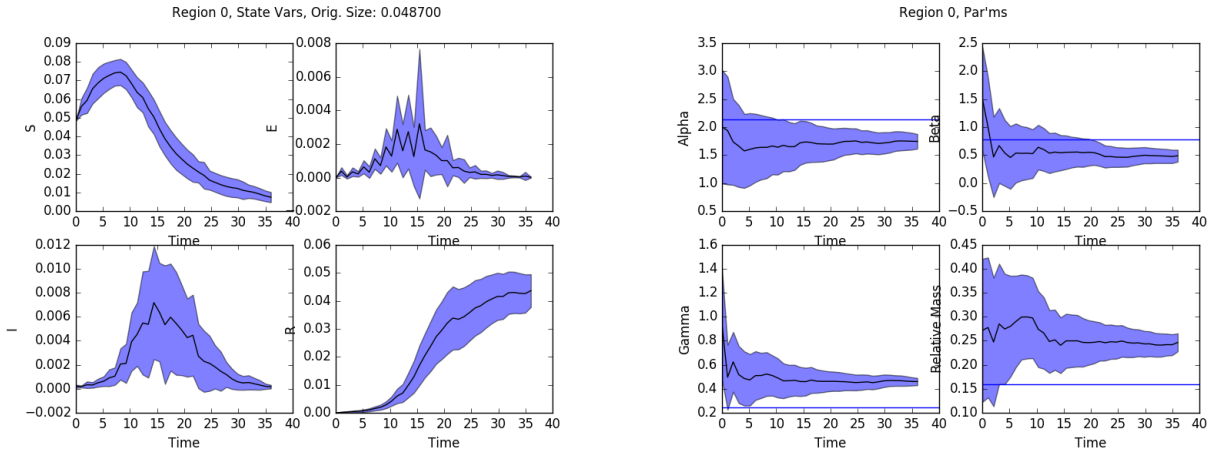


(a) (b)

Figure 1: (a) State estimates and (b) Parameter estimates for the '03-'04 Flu Season in HHS Region 1 (New England)

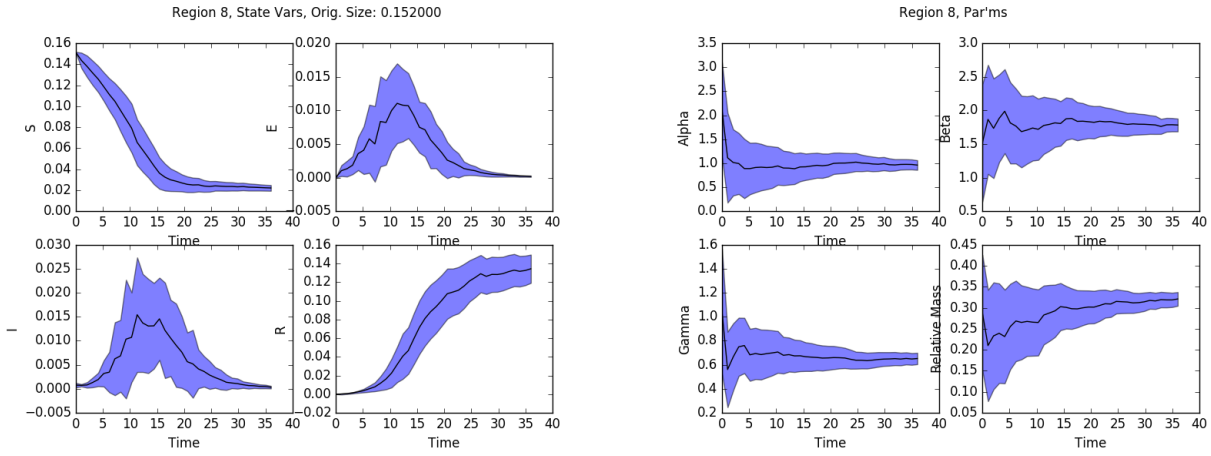
with 10, coupled regions). Fig. 2 gives the state and parameter estimates, with the true parameter values indicated by horizontal lines.

The synthetic data was generated by a continuous time epidemic model and the filter assumed a discrete time model, so it's possible that the incorrect parameter estimates (Fig. 2b) are due to model error, however this likely doesn't explain all of the parameter estimate error. I used the continuous time model to generate the synthetic data because the discrete time version does not respect the various problem constraints. In the dis-



(a) (b)

Figure 2: Inference from synthetic data for (a) State estimates, (b) Parameter estimates



(a) (b)

Figure 3: (a) State estimates and (b) Parameter estimates for the '03-'04 Flu Season in HHS Region 9 (CA, NV, AZ and some island territories)

crete time model the state variables often went negative or exceeded the bound of total population size.

The misbehavior of the discrete time model reflects similar misbehaviors in the filter. While it's shown in Fig. 1 plot, all parameter estimates were fairly unstable for between different filtering runs (for both the Flu Trends data and the synthetic data). The state estimates were more consistent across runs, but they are almost certainly incorrect. Fig. 1a shows that the filter is concluding that New England experienced a total epidemic size that was around twice its initial population size. Having lived in New England during this time period, I can confirm that this did not happen.

This trend was similar across the 10 regions, with epidemic sizes being unrealistically large. Furthermore the filter pulled each initial susceptible population towards around 1/10 of the total US population. In Fig. 1a we see that the susceptible population jumps upwards, if we compare this to the much larger Region 9 (Fig. 3a) we see that the susceptible population almost immediately decreases to around 1/10 of the total US population, at which point norm epidemic dynamics take over. I suspect that the filter is exploiting the coupling between regions to keep e_t at a size that can match the observed growth rate. This suggests that an alternative model for regional interactions might be better, possibly one that is not as symmetric.

Another issue encountered with the filter is that occasionally an ensemble member flips to negative state values, or it's total population size increases past the (fixed) US population. To mitigate this I had to choose fairly tight parameter priors, which may be contributing to the instability in the parameter estimates. A variational approach could allow these constraints to be factored into the filtering explicitly, and so may be more suitable for epidemic surveillance.

Even without these difficulties, I think there is a clear and overwhelming identifiability issue with this problem. There are probably several combinations of parameters which could produce similar infection growth rate time series, and I suspect that this is really the main cause of the parameter estimate instability. It seems reasonable to conclude that epidemic surveillance needs to include at least one more observation of either another state variable or a parameter to be really tractable. In my opinion the best candidate is the region coupling parameters, as this is something that could be estimated from daily flight information.

5 Conclusion

In this project we showed that the square root EnKF is not suitable for multi-region flu surveillance using Google Flu Trends. While the high dimensionality of the model was not a (direct) issue for this filter, both the lack of parameter identifiability (due to a limited observation model) and the constraints on the state variables and parameter values mean that any estimates produced by the filter are almost certainly incorrect. I found that for a synthetic time series the SRF failed to indentify the true parameter values (although this may have been due to model error). Given these issues, a data assimilation approach to flu surveillance will likely need a more sophisticated observation network, a variational filtering algorithm, and a better process model. These changes will hopefully improve parameter identifiability and deal with the problem constraints better.

References

[1] Dukic, Lopes, and Polson (2012). JASA, 107:500, 1410-1426.

[2] Tippett, Anderson, Bishop, Hamilton, and Whitaker (2003). MWR, 131, 1485–1490.

[3] Annan and Hargreaves (2004). Tellus, 56a, 520-526.

[4] Grooms (2016), Unpublished Course Notes