Influenza Trend Prediction Using Kalman Filter and Particle Filter

Ying Zhang
yingz1@andrew.cmu.edu
Department of Physics, Carnegie Mellon University

Committee: Prof. Roni Rosenfeld, Prof. Ryan Tibshirani, and Prof. Roy Maxion

Abstract

Background. Seasonal influenza can cause severe health problems and significant economic burdens in various regions of the world. In addition to the substantial morbidity and mortality cases caused by influenza, the emergency department crowding is also partially attributed to the influenza patients. Forecasting the influenza trends is crucial in developing effective countermeasures to mitigate an epidemic outbreak and allows medical facilities to allocate resources accordingly.

Aim. Filtering techniques are studied to model various dynamic systems, because they provide better estimations through recursive Bayesian updates. We review and implement several filtering techniques to predict the influenza trend in short term.

Data. We studied both the synthetic data generated from an epidemic mechanistic model, and real influenza data from three different sources. The synthetic data is based on the mechanistic differential equation (SIRS model) with added noise. The Centers for Disease Control and Prevention reports the incidences of influenza-like illness (ILI) through its surveillance network. Web data of Twitter messages and Wikipedia article access logs are shown to be highly correlated with the ILI data; therefore, the relative real-time web data allows the robust prediction of influenza when the ILI data is not available due to delay in the surveillance network.

Methods. Even though Kalman filters and particle filters are previously widely applied to engineering problems, the study of infectious diseases using filtering methods is a recent advancement. We first implement the filtering methods in combination with the mechanistic model to test the prediction ability using the synthetic data. The synthetic data allows a quantitative comparison based on the mean square error and the log likelihood for the different filtering methods. To study the real influenza trend, we implement the extended Kalman filter and particle filter using ILI data, Wikipedia and twitter signals in a recently developed empirical framework (Archetype framework).

Results. In the experiments of the synthetic data, unscented Kalman filter yields the lowest mean square error and the highest log likelihood in comparison with extended Kalman filter, ensemble Kalman filter and particle filter. The mean square error from the unscented Kalman filter is 5% smaller than the ensemble Kalman filter. In forecasting the influenza trend, the real influenza observations are well within 80% confidence interval of one-week predictions using the Archetype framework. However the influenza peak prediction is lagged by 1 week than the observed influenza peak.

Conclusions The filtering methods demonstrate fine performance in both the synthetic data and real influenza data. Filtering methods in the Archetype framework are simpler to implement, and yield good influenza predictions.

1 Introduction

Seasonal influenza is one of the significant causes of morbidity and mortality around the world. It reults in 250-500 thousand deaths annually world-wide, and contributes to approximately $87.1 billion in economic burden in the United State alone (WHO, 2009). In many countries, surveillance data for influenza-like illness has been documented, for example, the Centers for Disease Control and Prevention (CDC) has tracked the weekly influenza-like disease for the past three decades (CDC, 2012). A proper modeling for the influenza surveillance data is helpful for both improving health conditions and reducing economic burdens. Short-term forecasts within a season can help individuals and organizations adjust activity plans to reduce influenza transmission while long-term predictions are valuable for selecting vaccines for future seasons (Brooks, Farrow, Hyun, Tibshirani, & Rosenfeld, 2015).

Several mechanistic models (also called compartmental models) have been proposed to describe the epidemics based on the physics of disease propagation (Brauer, Castillo-Chavez, & Castillo-Chavez, 2001; Newman, 2002). For example, the SIRS (susceptible-infectious-recovered-susceptible) model describes the transition between population proportions which are susceptible to influenza, infected by the virus and recovered from the infection. Assumptions imposed by the SIRS model include fully mixed population and identical transmission behavior for different strains of influenza. Thus, current prediction ability based on mechanistic models alone for the timing or magnitude of influenza outbreaks is limited due to factors such as spatial heterogeneity, preferential mixing in the network structure for human interactions (Shaman & Karspeck, 2012).

Besides the mechanistic models, other influenza modeling approaches can mostly be categorized as agent-based models, parametric statistical models and empirical Bayes framework. The agent-based models rely on the complex interaction and disease pattern among the population, and are generally applied to special cases of a single strain of influenza (Colizza, Barrat, Barthelemy, Valleron, & Vespignani, 2007; Grefenstette et al., 2013). The parametric statistical models utilize various time series analysis methods to predict the influenza trend. Recent influenza prediction studies have used Box-Jenkins methods (Shumway & Stoffer, 2013), for example, autoregressive integrated moving average (ARIMA) model (Quenel & Dab, 1998; Soebiyanto, Adimi, & Kiang, 2010) and generalized autoregressive moving average (GARMA) model (Dugas et al., 2013). The GARMA forecast model integrated with Google Flu Trends information yields good prediction power even though the transmission mechanism was not considered (Dugas et al., 2013). The newly developed empirical Bayes framework does not make strong domain-specific assumptions, thus can be easily applied to some other diseases with seasonal epidemics (Brooks et al., 2015).

In the "real-world", the transmission of influenza is better depicted by stochastic difference models (at discrete times) in determining the susceptible or infectious population. In addition to the noise in the transmission process, there are always uncertainties associated with real-time measurements for the influenza counts, e.g. underestimate for the asymptomatic population, delayed reporting, etc.(Laporte, 1993). The feature of the uncertainty occurrence in both the process and the observation of the influenza dynamic system encourages the application of filtering techniques. Filtering techniques are known to provide better estimates of a dynamic system based on the measurements from the past to the current time through a recursive Bayesian update (Maybeck, 1982; Evensen, 2009).

2 Problem and approach

In this report, we first overview various filtering methods and examine their underlying assumptions and relative advantage in estimating the unobserved states. We implement the extended Kalman filter, unscented Kalman filter, ensemble Kalman filter and particle filter to the synthetic data generated by the SIRS model. The noisy synthetic data allows the evaluation of different filters, where the unobserved states are known.

To model the real influenza trend, we consider an empirical framework where the typical trend of influenza is depicted by an Archetype function with uncertainty in scale and time shift. We examine whether the filtering techniques in this empirical framework can provide a robust prediction of the influenza trend.

3 Background and related work

Filtering techniques are often utilized to study dynamic system problems in estimating the internal states (parameters or hidden variables) when the system is partially observable and random noise is present in both the dynamic process and the observations(Maybeck, 1982; Grewal & Andrews, 2014). Through a Bayesian update, the estimated system states can be improved when observational data are available on-line (Doucet, De Freitas, & Gordon, 2001). Significant portions of filters are applied in the Markovian state space (Doucet & Johansen, 2009).

Consider a discrete-time Markov process, where the hidden (unobserved) states are denoted by $\{\mathbf{x}_t : t \in \mathbb{N}\}$ with a initial distribution $p(\mathbf{x}_0)$ and the transitional probability $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. The observations $\{\mathbf{z}_t : t \in \mathbb{N}_+\}$ are assumed to be conditionally independent with each other given the states $\{\mathbf{x}_t\}$, and the conditional probability distribution is $p(\mathbf{z}_t|\mathbf{x}_t)$. In terms of equations, the model is fully specified by the prior, the transitional probability and the conditional probability (Doucet et al., 2001),

$$p(\mathbf{x}_0), \quad p(\mathbf{x}_k|\mathbf{x}_{k-1}) := f(\mathbf{x}_k|\mathbf{x}_{k-1}), \quad p(\mathbf{z}_k|\mathbf{x}_k) := g(\mathbf{z}_k|\mathbf{x}_k). \tag{1}$$

Using a simplified notation $\mathbf{x}_{i:j} := (\mathbf{x}_i, \mathbf{x}_{i+1}, ...\mathbf{x}_j)$, the marginal distribution for $\{\mathbf{x}_n\}_{n\geq 1}$ and the likelihood can be expressed as,

$$p(\mathbf{x}_{0:n}) = p(\mathbf{x}_0)\prod_{k=1}^{n} f(\mathbf{x}_k|\mathbf{x}_{k-1}), \quad p(\mathbf{z}_{1:n}|\mathbf{x}_{1:n}) = \prod_{k=1}^{n} g(\mathbf{z}_k|\mathbf{x}_k) \tag{2}$$

The task of inference is to estimate the states given the observations, i.e. $p(\mathbf{x}_j|\mathbf{z}_{1:k})$. Depending on the value of j, the inference problem can be categorized as 1) filtering if $j = k$, 2) smoothing if $j < k$ and 3) prediction if $j > k$ (Murphy, 2012). In terms of influenza prediction, the objective is to estimate the posterior distribution conditioning on all the past data, i.e. $p(\mathbf{x}_{k+1}|\mathbf{z}_{1:k})$. This can be achieved by combining the filtering estimate $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ with a transitional probability of $p(\mathbf{x}_{k+1}|\mathbf{x}_k)$.

Among the various filtering methods, Kalman filter is known to be the optimal linear filter to model systems with input and output, where the update of measurements can improve the prediction (Maybeck, 1982; Welch & Bishop, 2006). It is based on two assumptions regarding the noise in the processes and measurements, i.e. Gaussian noise and white noise. Relaxing the Gaussian noise assumption, Kalman filter is still the best filter out of the class of unbiased linear

filters (Welch & Bishop, 2006). When the noise is not white, i.e, there is autocorrelation between noise, a small linear system can be combined with the Kalman filter such that the noise for the whole system is white (Welch & Bishop, 2006; Grewal & Andrews, 2014). Because of these robust properties, Kalman filter has been widely applied in the engineering field and acclaimed as one of the most important algorithms invented in the 20th century (Casti, 2000).

The original Kalman filter is limited to linear dynamic systems. For nonlinear systems, the extended Kalman filter has been developed, which involves a first order Taylor expansion to approximate the nonlinear relations (Julier & Uhlmann, 2004). Both the Kalman filter and the extended Kalman filter rely on the update of mean and covariance matrix for the states, because Gaussian distributions are fully specified by the mean and covariance. Another extension of the Kalman filter is the unscented Kalman filter (UKF), which utilizes the property of sigma points (a minimal set of carefully chosen sample points to represent a Gaussian random variable) to improve the approximation accuracy (Julier & Uhlmann, 1997). It is proven that the sigma points can capture the mean and covariance of the non-linearly transformed Gaussian random variable with third order accuracy in terms of Taylor expansion. Interestingly, the UKF possesses the same order of computational complexity as that of the EKF while improving the approximation accuracy (E. Wan & Van Der Merwe, 2000).

For systems with high dimensions, the manipulation of covariance matrix can be computationally expensive. Therefore, the ensemble Kalman filter (EnKF) is proposed in literature to reduce the computational cost (Mandel, 2009). The EnKF uses a random sample, an ensemble, to represent the distribution of the system such that, the updating of probability distribution is achieved by updating the members of the ensemble. The ensemble approach overcomes the high computational cost of maintaining the covariance matrix at high dimensions. Despite the low computational cost, EnKF method still assumes the noise generated from the process and the measurement is Gaussian, which limits the application to nonlinear problems (Mandel, 2009; Gillijns et al., 2006).

In contrast to the Kalman filters, the particle filter method does not assume Gaussian random variables; therefore the later can be widely applied to nonlinear and non-Gaussian processes. The formulation of particle filters is similar to the ensemble Kalman filter in terms of utilizing samples to approximate the distribution (Mandel, 2009). Based on the sequential Monte Carlo method, particle filters approximate the posterior distribution with updated observations. The flexibility of particle filters in modeling general situations is accompanied by the drawback of higher computational cost compared with Kalman filters (Doucet et al., 2001).

Despite the wide applications in engineering fields, the adaptation of filtering methods to model infectious disease is a relatively recent advancement. Ionides et al. presented the iterated filtering method, which could achieve a maximum likelihood estimate for partially-observed nonlinear stochastic dynamic systems (Ionides, Bretó, & King, 2006). The application of filtering methods to the cholera study improved the epidemic simulation, using the mortality measurements collected at various regions in different years (King, Ionides, Pascual, & Bouma, 2008). One additional benefit of the filtering process is the dual estimation of the state variables and the dynamic parameters (E. A. Wan, Van Der Merwe, & Nelson, 2000).

Shaman et al. applied multiple filtering techniques in combination with mechanistic models to predict the influenza trend (Shaman, Pitzer, Viboud, Grenfell, & Lipsitch, 2010; Shaman & Karspeck, 2012; Yang, Karspeck, & Shaman, 2014; Yang, Lipsitch, & Shaman, 2015). The four mechanistic models examined are SIR, SIRS, SEIR (susceptible-exposed-infectious-recovered) and SE^2I^2R (susceptible-exposed, stage 1-exposed, stage 2-infectious,stage 1-infectious, stage 2-

recovered). Both the particle filter and the ensemble Kalman filter are implemented. The comparison suggests that SIRS model in combination with filtering methods yields the most reliable prediction. Accuracies for the particle filter and the ensemble Kalman filter are comparable in predicting the influenza activity for current week. In contrast, the particle filter performs slightly better in predicting peaks in near future (1-5 weeks) while the ensemble Kalman filter is more accurate for the peaks in the past.

4 Data

4.1 Synthetic data generated by the SIRS model

4.1.1 A mechanistic model: SIRS model

Figure 1 shows the transition of different population in the SIRS model: susceptible population can be infected while infectious population can recover from the flu, and finally recovered population may again be susceptible to the flu (Newman, 2002). In this model, S is the susceptible population, I is the infectious population, N is the total population, and $N-I-S$ is the recovered population; L is the average duration of immunity, D is the mean infectious period, β is the contact rate, and t is the time.

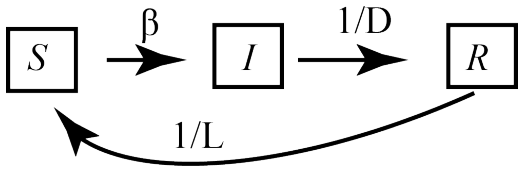

Figure 1: The SIRS model for the transmission of influenza among different class of subjects, where S, I and R represent susceptible, infectious and recovered population, respectively. The sum of S, I and R is the total population N.

In terms of differential equations, the SIRS model is governed by,

$$\begin{aligned} \frac{dS}{dt} &= \frac{N-S-I}{L} - \frac{\beta IS}{N}, \\ \frac{dI}{dt} &= \frac{\beta IS}{N} - \frac{I}{D}, \end{aligned} \tag{3}$$

where the conversion rate of susceptible population into infectious population is proportional to both the current susceptible population and infectious population, while inversely proportional to the total population, i.e. $\frac{\beta IS}{N}$. The rate of recovered population converting into susceptible population is inversely proportional to the duration of immunity while proportional to the recovered population, i.e. $\frac{N-S-I}{L}$. Finally the transition rate of infectious population to recovered population is $\frac{I}{D}$, inversely proportional to the mean infectious period.

4.1.2 Procedure for generating synthetic data

The differential equations in the SIRS (Equation 3) can be rewritten as difference equations at different days k. Define the state variable as $\mathbf{x}_k^\top = (S_k, I_k)$, i.e. the susceptible and infectious

population at different days. Denote the observed variable as z_k, which is the measured infected population. The process equations are set to be,

$$\begin{aligned} S_k &= S_{k-1} + \frac{N - S_{k-1} - I_{k-1}}{L} - \frac{\beta I_{k-1} S_{k-1}}{N} + Q_1 \\ I_k &= I_{k-1} + \frac{\beta I_{k-1} S_{k-1}}{N} - \frac{I}{D} + Q_2. \end{aligned} \tag{4}$$

with Q_1 and Q_2 being process noise. The measurement equation has the measurement noise R,

$$z_k = I_k + R. \tag{5}$$

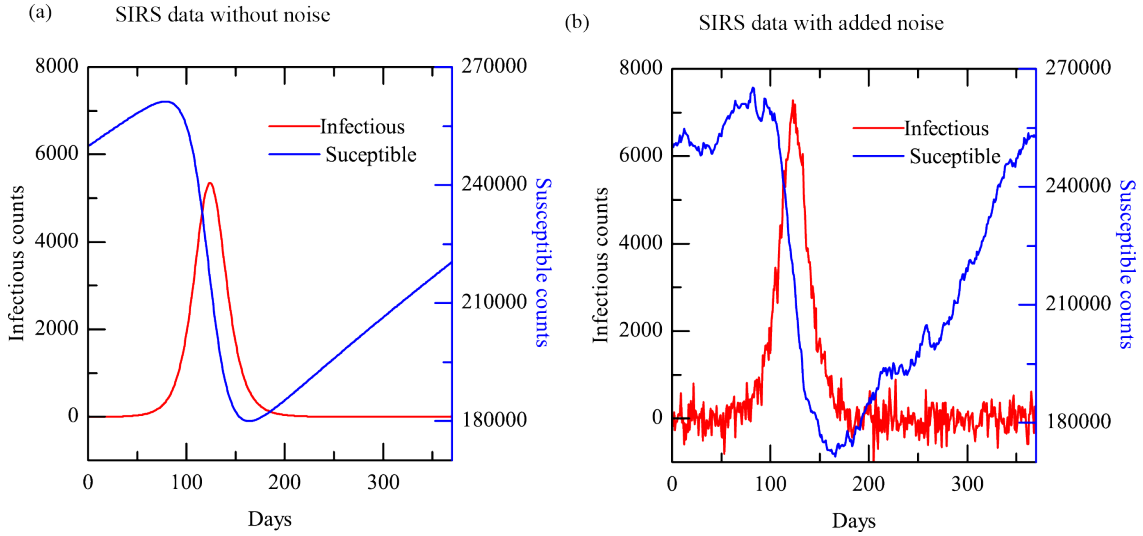

Figure 2: (a) The SIRS data generated by the difference equations (Equation 4) without any added noise. The blue curve represents the susceptible population while the red curve is the infected population. The peak of infected population is lagged by the peak of susceptible population (b) The synthetic SIRS data generated by the difference equations (Euqation 4) with added noise.The peak value for the infected population is larger than that in part (a) because of the random noise.

Figure 2a shows a realization of the SIRS model without any added noise. The parameters are set to roughly mimic the transmission properties of the influenza trend in New York of the year 1972: $\beta = 1; D = 2.3; L = 1425$ (Shaman & Karspeck, 2012). And the population is set artificially to be $N = 500,000, S_0 = 250,000, I_0 = 1$. Figure 2a shows that there is a single peak for infectious people which is lagged by the peak of susceptible population. This simple realization of SIRS model shows the pattern of influenza onset, peak and decline.

Figure 2b shows the synthetic data from the SIRS model with added noise. The process noise is set to follow Gaussian distribution, $Q_1 \sim N(0, 1000^2)$, $Q_2 \sim N(0, 0.1^2)$, and the measurement noise is also Gaussian $R \sim N(0, 300^2)$. The process equation and the observation equation specify both the conditional probability and the likelihood function (Equation 2). Both the susceptible and infected population demonstrate some deviation from the smooth version without noise. In Figure 2b, the peak value of infected population is higher than Figure 2a because of the random noise, while the general trends in the two figures are similar.

4.2 Real influenza data from CDC, Twitter messages and Wikipedia access logs

The CDC provides surveillance data for influenza-like illness (ILI) in the United States (CDC, 2012). The CDC compiles the data provided by the U.S. Outpatient Influenza-like Illness Surveillance Network who voluntarily reports the total patient visits and ILI visits. The data only counts the incidence of ILI, because doctors do not generally differentiate influenza from similar symptoms. Nonetheless, the ILI data gives a good trend for the influenza transmission (Shaman & Karspeck, 2012).

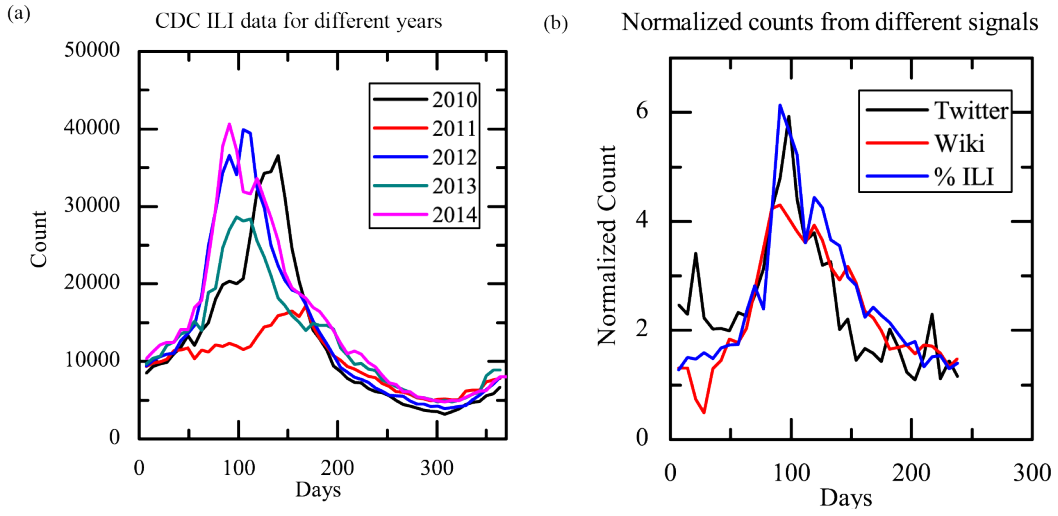

Figure 3: Visualization of influenza data (a) CDC reported patients visits due to influenza like-illness (ILI) in the United States (b) Normalized influenza data from Twitter, Wikipedia and ILI for year 2015.

Figure 3a shows the visualization of ILI counts each year from 2010 to 2014. Infected population typically has a peak around Day 100, but can shift in different years. Within one year, infected population can have several shallow peaks.

Though the ILI data provides a general trend of the influenza infections, it only counts the patients who seek for medical help, whereas there are significant portions of infected people who do not visit medical facilities. Also due to the bureaucratic hierarchy of the surveillance system, data availability may be lagged by 1-2 weeks (Hickmann et al., 2015). It is shown that Wikipedia article access logs and Twitter messages can be used as supplemental information to the ILI data with real time properties (Hickmann et al., 2015; Paul, Dredze, & Broniatowski, n.d.).

Figure 3b shows the normalized Twitter signals, Wikipedia access counts and ILI data in the year 2015 compiled by David Farrow (Farrow, 2016). The peaks from three different sources agree with each other with one to two weeks in difference. The general trends of the three data series also agree with high correlations.

5 Experimental methods

5.1 Filtering methods

5.1.1 Kalman filter

Let $\mathbf{x} \in \mathbb{R}^n$ be the state vector, and $\mathbf{z} \in \mathbb{R}^m$ be the observation vector. The Kalman filter considers a discrete-time linear system (Welch & Bishop, 2006),

$$\mathbf{x}_k = \mathbf{F}_k \mathbf{x}_{k-1} + \mathbf{w}_{k-1} \quad \text{(process)}, \qquad \mathbf{z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k \quad \text{(observation)}. \tag{6}$$

where $\mathbf{F}_k$ is the state transition operator applied to the previous state, and $\mathbf{H}_k$ is the observation operator mapping the state space to the observation space. $\mathbf{w}_k$ and $\mathbf{v}_k$ are the process and the observation noise respectively, and both assumed to be Gaussian, $\mathbf{w}_k \sim N(0, \mathbf{Q}_k)$, $\mathbf{v}_k \sim N(0, \mathbf{R}_k)$. Because both the process and observation are linear, if initial state $\mathbf{x}_0$ is Gaussian, then all the subsequent states $\mathbf{x}_k$'s and observations $\mathbf{z}_k$'s are also Gaussian.

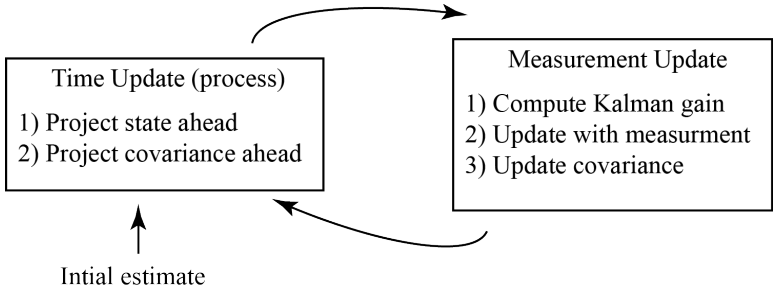

Figure 4: Kalman filter at each iteration includes a prediction step (based on the process equation) and an update step (using observations).

Figure 4 is a schematic of the Kalman filter, which uses sequential observations to estimate the system states. It is performed in two steps recursively: in the prediction step, the previous state is used to predict current state; in the update step, current observation is used to improve the state estimate. Each iteration involves an update of the mean and the covariance matrix.

Following the convention used in (Gillijns et al., 2006), we use subscript "f" to denote a priori estimate of state in the prediction step, and use "a" to denote a posterior estimate of the state with measurement update. The state variable estimate $\mathbf{x}_k^{\mathrm{f}}$ and a priori estimate error covariance $\mathbf{P}_k^{\mathrm{f}}$ are,

$$\begin{aligned} \mathbf{x}_k^{\mathrm{f}} &= \mathbf{F}_{k-1} \mathbf{x}_{k-1}^{\mathrm{a}} \\ \mathbf{P}_k^{\mathrm{f}} &= \mathbf{F}_{k-1} \mathbf{P}_{k-1}^{\mathrm{a}} \mathbf{F}_{k-1}^{\top} + \mathbf{Q}_{k-1} \end{aligned} \tag{7}$$

where $\mathbf{P}_k^{\mathrm{f}} = \mathbb{E}[(\mathbf{x}_k - \mathbf{x}_k^{\mathrm{f}})(\mathbf{x}_k - \mathbf{x}_k^{\mathrm{f}})^{\top}]$.

In the update step, the Kalman gain $\mathbf{K}_k$, the posterior estimate of the state $\mathbf{x}_k^{\mathrm{a}}$ and the posterior covariance $\mathbf{P}_k^{\mathrm{a}}$ are,

$$\begin{aligned} \mathbf{K}_k &= \mathbf{P}_{k,xz}^{\mathrm{f}} (\mathbf{P}_{k,zz}^{\mathrm{f}})^{-1}, \\ \mathbf{P}_k^{\mathrm{a}} &= (I - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_k^{\mathrm{f}}, \\ \mathbf{x}_k^{\mathrm{a}} &= \mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k (\mathbf{z}_k - \mathbf{H}_k \mathbf{x}_k^{\mathrm{f}}) \end{aligned} \tag{8}$$

where the intermediate steps for the Kalman gain involves two covariance matrix,

$$\begin{aligned}\mathbf{P}_{k,xz} &= \mathbb{E}[(\mathbf{x}_k - \mathbf{x}_k^{\mathrm{f}})(\mathbf{z}_k - \mathbf{H}_k\mathbf{x}_k^{\mathrm{f}})^\top] = \mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^\top \\ \mathbf{P}_{k,zz} &= \mathbb{E}[(\mathbf{z}_k - \mathbf{H}_k\mathbf{x}_k^{\mathrm{f}})(\mathbf{z}_k - \mathbf{H}_k\mathbf{x}_k^{\mathrm{f}})^\top] = \mathbf{H}_k\mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^\top + \mathbf{R}_k\end{aligned} \tag{9}$$

5.1.2 Extended Kalman filter

Consider the nonlinear extension of the process and observation equations (Equation 6).

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}) + \mathbf{w}_{k-1} \quad \text{(process)}, \qquad \mathbf{z}_k = \mathbf{h}(\mathbf{x}_k) + \mathbf{v}_k \quad \text{(observation)}. \tag{10}$$

The extended Kalman filter uses the linearization based on Taylor expansion,

$$\mathbf{F}_k = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{x}_k^{\mathrm{a}}}, \quad \mathbf{H}_k = \frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{x}_k^{\mathrm{a}}} \tag{11}$$

The original Kalman filter formulation (Equation 7 and Equation 8) can be rewritten as an extended Kalman filter with prediction step,

$$\begin{aligned}\mathbf{x}_k^{\mathrm{f}} &= \mathbf{f}(\mathbf{x}_{k-1}) \\ \mathbf{P}_k^{\mathrm{f}} &= \mathbf{F}_{k-1}\mathbf{P}_{k-1}^{\mathrm{a}}\mathbf{F}_{k-1}^\top + \mathbf{Q}_{k-1}\end{aligned} \tag{12}$$

and the update step:

$$\begin{aligned}\mathbf{K}_k &= \mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^\top(\mathbf{H}_k\mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^\top + \mathbf{R}_k)^{-1}, \\ \mathbf{P}_k^{\mathrm{a}} &= (I - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^{\mathrm{f}}, \\ \mathbf{x}_k^{\mathrm{a}} &= \mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k(\mathbf{z}_k - \mathbf{h}(\mathbf{x}_k^{\mathrm{f}}))\end{aligned} \tag{13}$$

5.1.3 Unscented Kalman filter

Unscented Kalman filter relies on the unscented transformation, which is used for determining the statistics of a random variable through a nonlinear transformation. For an L-dimensional random variable $\mathbf{x}$ through a nonlinear function $\mathbf{y} = \mathbf{f}(\mathbf{x})$, the statistics of $\mathbf{y}$ can be determined using the following procedures (E. Wan & Van Der Merwe, 2000). Assume $\mathbf{x}$ has mean $\bar{\mathbf{x}}$ and covariance $\mathbf{P}_{\mathbf{x}}$, $2L+1$ sigma vectors $\mathcal{X}_i$ are determined as,

$$\begin{aligned}\mathcal{X}^0 &= \bar{\mathbf{x}} \\ \mathcal{X}^i &= \bar{\mathbf{x}} + \left(\sqrt{(L+\lambda)\mathbf{P}_{\mathbf{x}}}\right)_i \qquad i = 1, \ldots, L \\ \mathcal{X}^i &= \bar{\mathbf{x}} - \left(\sqrt{(L+\lambda)\mathbf{P}_{\mathbf{x}}}\right)_{i-L} \qquad i = L+1, \ldots, 2L \\ W_m^0 &= \lambda/(L+\lambda) \\ W_c^0 &= \lambda/(L+\lambda) + (1 - \alpha^2 + \beta) \\ W_m^j &= W_c^j = 1/[2(L+\lambda)] \qquad j = 1, \ldots, 2L\end{aligned}$$

where $\lambda = \alpha^2(L+\kappa) - L$ is scaled by α and κ. α is typically set as a small parameter (e.g. 0.001) which represents the spread of sigma points around the mean. κ is usually set as $\kappa = 0$ while the other parameter is set to be $\beta = 2$ for Gaussian distribution of $\mathbf{x}$. And $\left(\sqrt{(L+\lambda)\mathbf{P}_{\mathbf{x}}}\right)_i$ represents the i-th row of the square root matrix.

Nonlinear transformation of the sigma points yields,

$$\mathcal{Y}^i = \mathbf{f}(\mathcal{X}^i), \qquad i = 0, \ldots, 2L+1. \tag{14}$$

The mean and covariance of $\mathbf{y}$ are then approximated by

$$\bar{\mathbf{y}} \approx \sum_{i=0}^{2L} W_m^i \mathcal{Y}^i \tag{15}$$

$$\mathbf{P}_\mathbf{y} \approx \sum_{i=0}^{2L} W_c^i \left(\mathcal{Y}^i - \bar{\mathbf{y}}\right)\left(\mathcal{Y}^i - \bar{\mathbf{y}}\right)^\top \tag{16}$$

To apply the unscented transformation to Kalman filter, these procedures would correspond to the prediction step, with $\bar{\mathbf{x}} \to \mathbf{x}_{k-1}^{\mathrm{a}}$, $\mathbf{P}_\mathbf{x} \to \mathbf{P}_{k-1}^{\mathrm{a}}$, $\bar{\mathbf{y}} \to \mathbf{x}_k^{\mathrm{f}}$, and $\mathbf{P}_\mathbf{y} \to \mathbf{P}_k^{\mathrm{f}}$. And in the measurement step the covariance is $\mathbf{P}_k^{\mathrm{f}} \to \mathbf{P}_k^{\mathrm{a}}$, since it is used for resampling $\left(\sqrt{(L+\lambda)\mathbf{P}_k^{\mathrm{a}}}\right)$ in the next iteration.

5.1.4 Ensemble Kalman filter

For high-dimensional systems, ensemble Kalman filter (EnKF) is proposed to reduce the computational cost in manipulating matrices. Ensemble Kalman filter utilizes a group of particles (ensemble) to represent the distribution of the system, and replace the covariance matrix by the sample covariance of the ensemble(Mandel, 2009). Therefore, the evolution of the distribution is achieved by updating each member of the ensemble. At time $k-1$, given an ensemble of n state estimates, $\{\mathbf{x}_{k-1}^{\mathrm{a},i}\}$, $i = 1, \ldots, n$, the prediction step yields an ensemble of states $\mathbf{x}_k^{\mathrm{f}}$, and priori state sample covariance $\hat{\mathbf{P}}_k^{\mathrm{f}}$

$$\mathbf{x}_k^{\mathrm{f},i} = \mathbf{f}(\mathbf{x}_{k-1}^{\mathrm{a},i}) + \mathbf{w}_{k-1}^i, \tag{17}$$

$$\hat{\mathbf{P}}_k^{\mathrm{f}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_k^{\mathrm{f},i} - \overline{\mathbf{x}_k^{\mathrm{f},i}}\right)\left(\mathbf{x}_k^{\mathrm{f},i} - \overline{\mathbf{x}_k^{\mathrm{f},i}}\right)^\top \tag{18}$$

where $\mathbf{w}_{k-1}^i \sim N(0, \mathbf{Q}_{k-1})$ and $\overline{\mathbf{x}_k^{\mathrm{f},i}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_k^{\mathrm{f},i}$ is the ensemble mean.

Replacing the error covariance matrix in Equation 9, the sample error covariances for EnKF are $\hat{\mathbf{P}}_{k,xz}^{\mathrm{f}}$ between the predicted state and predicted observation, and $\hat{\mathbf{P}}_{k,zz}^{\mathrm{f}}$ for the predicted observation,

$$\begin{aligned}
\hat{\mathbf{P}}_{k,xz}^{\mathrm{f}} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_k^{\mathrm{f},i} - \overline{\mathbf{x}_k^{\mathrm{f},i}}\right)\left(\mathbf{h}(\mathbf{x}_k^{\mathrm{f},i}) - \overline{\mathbf{h}(\mathbf{x}_k^{\mathrm{f},i})}\right)^\top \\
\hat{\mathbf{P}}_{k,zz}^{\mathrm{f}} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{h}(\mathbf{x}_k^{\mathrm{f},i}) - \overline{\mathbf{h}(\mathbf{x}_k^{\mathrm{f},i})}\right)\left(\mathbf{h}(\mathbf{x}_k^{\mathrm{f},i}) - \overline{\mathbf{h}(\mathbf{x}_k^{\mathrm{f},i})}\right)^\top
\end{aligned} \tag{19}$$

where $\overline{\mathbf{h}(\mathbf{x}_k^{\mathrm{f},i})} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}(\mathbf{x}_k^{\mathrm{f},i})$.

In the update step, the Kalman gain $\hat{\mathbf{K}}_k$, the updated ensemble member $\mathbf{x}_k^{\mathrm{a},i}$ and the posterior

sample covariance $\hat{\mathbf{P}}_k^{\mathrm{a}}$ are,

$$\hat{\mathbf{K}}_k = \hat{\mathbf{P}}_{k,xz}^{\mathrm{f}} \left(\hat{\mathbf{P}}_{k,zz}^{\mathrm{f}} \right)^{-1} \tag{20}$$

$$\mathbf{x}_k^{\mathrm{a},i} = \mathbf{x}_k^{\mathrm{f},i} + \hat{\mathbf{K}} \left(\mathbf{z}_k + \mathbf{v}_k^i - \mathbf{h}(\mathbf{x}_k^{\mathrm{f},i}) \right) \tag{21}$$

$$\hat{\mathbf{P}}_k^{\mathrm{a}} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\mathbf{x}_k^{\mathrm{a},i} - \overline{\mathbf{x}_k^{\mathrm{a},i}} \right) \left(\mathbf{x}_k^{\mathrm{a},i} - \overline{\mathbf{x}_k^{\mathrm{a},i}} \right)^{\top} \tag{22}$$

where $\mathbf{v}_k^i \sim N(0, \mathbf{R}_k)$ is the random observation noise, such that $\mathbf{z}_k^i = \mathbf{z}_k + \mathbf{v}_k^i$ represents the perturbed observation. The posterior state estimate is the ensemble mean of the updated estimate. $\overline{\mathbf{x}_k^{\mathrm{a},i}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_k^{\mathrm{a},i}$

5.1.5 Particle filter

As explained by Arulampalam et al. (Arulampalam, Maskell, Gordon, & Clapp, 2002), particles can be used with different weights to approximate a distribution, which is not computationally expensive in low dimensions. Denote $\{\mathbf{x}_{0:k}^i, w_k^i\}$ as a set of particles to characterize the posterior distribution $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$, where $\{\mathbf{x}_{0:k}^i, i = 1, 2, ...N_s\}$ is a set of support points with weight $\{w_k^i\}$, then the posterior distribution is approximated by,

$$p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_{0:k} - \mathbf{x}_{0:k}^i). \tag{23}$$

The weights for particles are chosen based on the importance sampling principle which relies on the following condition. If $p(x) \propto \pi(x)$ is a probability distribution which is difficult to sample from but the form of $\pi(x)$ is analytically known, and another proposal distribution $q(x)$ is easy to sample from, then the weighted approximation to the density $p(x)$ is given by(Doucet et al., 2001),

$$p(x) \approx \sum_{i=1}^{N_s} w^i \delta(x - x^i), \quad \text{where } w^i \propto \frac{\pi(x^i)}{q(x^i)}.$$

Through the update of different weights of particles, the posterior distribution is therefore approximately updated at different time steps k. If the samples $\mathbf{x}_{0:k}$ were drawn from an importance density $q(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$, then the weights are then

$$w_k^i \propto \frac{p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})}{q(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})}$$

If we invoke the Markov property $q(\mathbf{x}_k|\mathbf{x}_{0:k-1}, \mathbf{z}_{1:k}) = q(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_k)$, then the importance density is only dependent on $\mathbf{x}_{k-1}$ and $\mathbf{z}_k$. In the common case where only a filtered estimate of $p(\mathbf{x}_k|\mathbf{z}_{1:k}$ is required, we only need to store $\mathbf{x}_k^i$ rather than the whole path $x_{0:k-1}^i$ and the history observations of $\mathbf{z}_{1:k-1}$.

It has been shown (Doucet et al., 2001; Doucet & Johansen, 2009; Arulampalam et al., 2002) that to approximate the posterior density $p(\mathbf{x}_k|\mathbf{z}_{1:k})$, the modified weight from time step $k-1$ to k is,

$$w_k^i \propto w_{k-1}^i \frac{p(\mathbf{z}_k|\mathbf{x}_k^i) p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i, \mathbf{z}_k)} \tag{24}$$

F where $q(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i, \mathbf{z}_k)$ is the new importance proposal distribution to sample from. A common choice for the importance sampling is using the prior distribution, i.e.

$$q(\mathbf{x}_k|\mathbf{x}_{k-1}^i, \mathbf{z}_k) = p(\mathbf{x}_k|\mathbf{x}_{k-1}^i) \tag{25}$$

This choice yields a simplified equation for weight update(Equation 24) at each time update,

$$w_k^i \propto w_{k-1}^i p(\mathbf{z}_k|\mathbf{x}_k^i) \tag{26}$$

The filtered posterior density $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ (different from posterior density $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$) is given by,

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i) \tag{27}$$

However, there is a degeneracy problem associated with the particle filter using the above weight update procedure. After finite iterations, all except one particle will have negligible weight, which does not yield an appropriate approximation for the posterior distribution(Arulampalam et al., 2002). To reduce the effect of degeneracy, a resampling step is necessary. In the resampling step, a new set of particles $\{\mathbf{x}_k^{i*}, w_k^{i*}\}$ are obtained by resampling with replacement N_s times from the approximate filtered posterior distribution $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ in Equation 27. This is indeed the bootstrap sampling from the discrete density of Equation 27, where each particle has the new weight $w_k^{i*} = 1/N$.

5.2 Filtering methods in the Archetype framework

The Archetype framework was developed by David Farrow (Farrow, 2016). Similar to the Empirical Bayes framework(Brooks et al., 2015), the Archetype framework also assumed that future epidemics will resemble the shape of past epidemics. The Archetype function is based on the canonical shape of an influenza epidemic, where the detailed procedures are described in the technical report(Farrow, 2016). The main components of the procedures are: 1) remove irregularities during holiday weeks, 2) smooth historical ILI data using Gaussian kernel smoother, 3)align historical peaks to the center of the season and 4)interpolate between different yearly curves and obtain the final Archetype function (shape).

Figure 5 shows the Archetype function: inputting a time in unit of week returns a value of the normalized infectious population. With the Archetype function, current seasonal influenza trend behavior can be modeled by two parameters: time shift and magnitude scale. Intuitively, the time shift depicts that future influenza counts are dependent on the current infectious population with uncertainty in time advancement or delay. And the magnitude scale can be interpreted as fluctuation in infected population. Filtering methods in the Archetype framework can also provide a better forecast of influenza trend when new data is available online through the recursive Bayesian update.

In terms of mathematical formulations, the Archetype framework can be described as a state space model. The states are represented by $\mathbf{x}_w$, and the observables are denoted by $\mathbf{z}_w$, where the subscript w denotes the unit of time, in weeks with values of [0, 51] inclusive. The definition of $w = 0$ corresponds to the calendar Week 30 (i.e. end of July, middle of influenza off season).

The state variable is a two dimensional vector which varies with weeks, $\mathbf{x}_w^\top = (t_w, s_w)$, where $t_w \in (-26, +26)$ is the shift parameter and $s_w \in \mathcal{R}_+$ is the scale parameter. Both parameters s_w

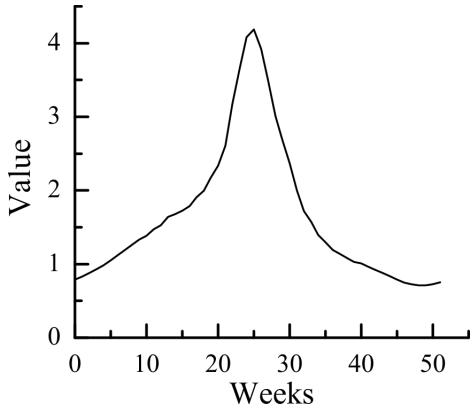

Figure 5: Archetype function to read out measurement

and t_w are dependent on the week w, and can be used to obtain the influenza counts from the Archetype function. The state space model is fully specified by the prior distribution, the process equation and the observation equation.

The prior distribution for the first week is assumed to be decomposable for s_1 and t_1, i.e.

$$P(\mathbf{x}_1) = P(s_1, t_1) = P(s_1)P(t_1) \tag{28}$$

where $s_1 \sim N_+(1, 0.1^2)$ is drawn from a truncated normal distribution such that s_1 is always positive, and t_1 is drawn from a normal distribution $t_1 \sim N(0, 3^2)$

The process equation for the two components of t_w and s_w follows,

$$t_w = t_{w-1} + q_t \tag{29}$$
$$s_w = \alpha s_{w-1} + (1 - \alpha) + q_s \tag{30}$$

where, q_t and q_s denote random variables from Gaussian distribution $q_t \sim N(0, \sigma_t^2), q_w \sim N(0, \sigma_s^2)$. $\alpha = 0.95$ is a parameter and $\sigma_t = \sigma_s = 0.2$. Or in the matrix form,

$$\mathbf{x}_w = \begin{pmatrix} 1 & 0 \\ 0 & \alpha \end{pmatrix} \mathbf{x}_{w-1} + \begin{pmatrix} 0 \\ 1 - \alpha \end{pmatrix} + \mathbf{q} \tag{31}$$

where $\mathbf{q}$ is a noise vector which is drawn from a Gaussian noise distribution:$\mathbf{q} \sim N(\mathbf{0}, Q)$ where Q denotes covariance matrix for the process noise, i.e.

$$Q = \begin{pmatrix} \sigma_t^2 & 0 \\ 0 & \sigma_s^2 \end{pmatrix}$$

The observation equation depends on the Archetype function $A(y)$, where y is defined in the range of [0, 52). We define an observation function h which acts on the state $\mathbf{x}$ without noise and holiday effects can be expressed as $h(\mathbf{x}_w) = s_w \cdot A[(w - t_w) \text{mod } 52]$.

Observations are three dimensional vectors (twitter signal, wiki signal, weighted ILI signal %wILI). Holiday affects the observation counts of %wILI, which is typically associated with the calendar time year end(Week 50, 51, 0, and Week 1), which corresponds to $w = 20, 21, 22, 23$.

therefore a multiplicative constant K_w is necessary to improve the observation.

$$K_w = \begin{cases} 1 & \text{if } w = 0, \ldots, 18, 19, 24, \ldots 51 \\ K_{20} & \text{if } w = 20 \\ K_{21} & \text{if } w = 21 \\ K_{22} & \text{if } w = 22 \\ K_{23} & \text{if } w = 23 \end{cases} \tag{32}$$

Denote the multiplicative vector as $\boldsymbol{\beta}$, which can be expressed as $\boldsymbol{\beta}^\top = (1, 1, K_w)$. Then the observation function is

$$\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}_w = s_w \cdot A[(w - t_w) \text{mod } 52] \begin{pmatrix} 1 \\ 1 \\ K_w \end{pmatrix} + \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix} \tag{33}$$

where $\mathbf{r}^\top = (r_1, r_2, r_3)$ is a noise vector with Gaussian distribution $\mathbf{r} \sim N(\mathbf{0}, R)$, where R denotes the covariance matrix for the measurement noise, i.e.

$$R = \begin{pmatrix} \sigma_{r1}^2 & 0 & 0 \\ 0 & \sigma_{r2}^2 & 0 \\ 0 & 0 & \sigma_{r3}^2 \end{pmatrix} \tag{34}$$

where $\sigma_{r1} = 0.7$ for twitter, $\sigma_{r2} = 0.5$ for wiki and $\sigma_{r3} = 0.5$ for unstable ILI. The dynamic process equations are then fully specified.

6 Experimental results

6.1 SIRS-Filter results for the synthetic data

The extended Kalman filter, ensemble Kalman filter, unscented Kalman filter and particle filter are implemented to estimate the state variables based on the measurements. The synthetic data allows a quantitative comparison of the filters in estimating the state variables since true values are known in advance. The initial estimate of the susceptible can impact the filtering results, thus we test the algorithms also on different initial values. More specifically, the filtering methods are run at two different prior values for the susceptible population: 250,500 (half a standard deviation away from the true value) and 500,000 (twice the true value).

Figure 6 shows the population estimates using the extended Kalman filter(EKF) and the unscented Kalman filter(UKF), where the initial susceptible population is close to the truth. In Figure 6a, both the susceptible population estimates agree with the true value in general while the EKF result is slightly more wiggly near the minimum. Figure 6b shows that the infectious population estimates overlap with each other and are quite smooth despite the noisy observations.

Figure 7 shows the population estimates using the particle filter(PF) and the ensemble Kalman filter(EnKF), where the initial susceptible value is close to the truth. In Figure 7a, the PF estimate for the susceptible population is slightly closer to the true value than the EnKF estimate. Figure 7b shows that the infectious population estimates again overlap with each other and are relatively smooth in spite of the noisy observations.

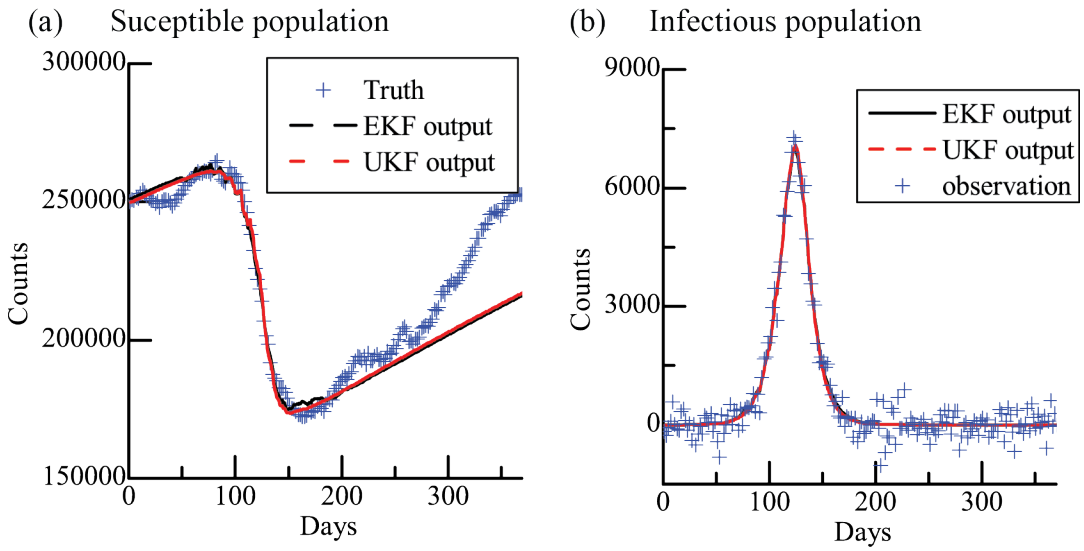

Figure 6: Implementation results of EKF and UKF where the initial estimate for the susceptible population is close to the true value. (a) Susceptible population for the synthetic truth and filtered estimates (b) Infectious population for the synthetic truth and filtered estimates

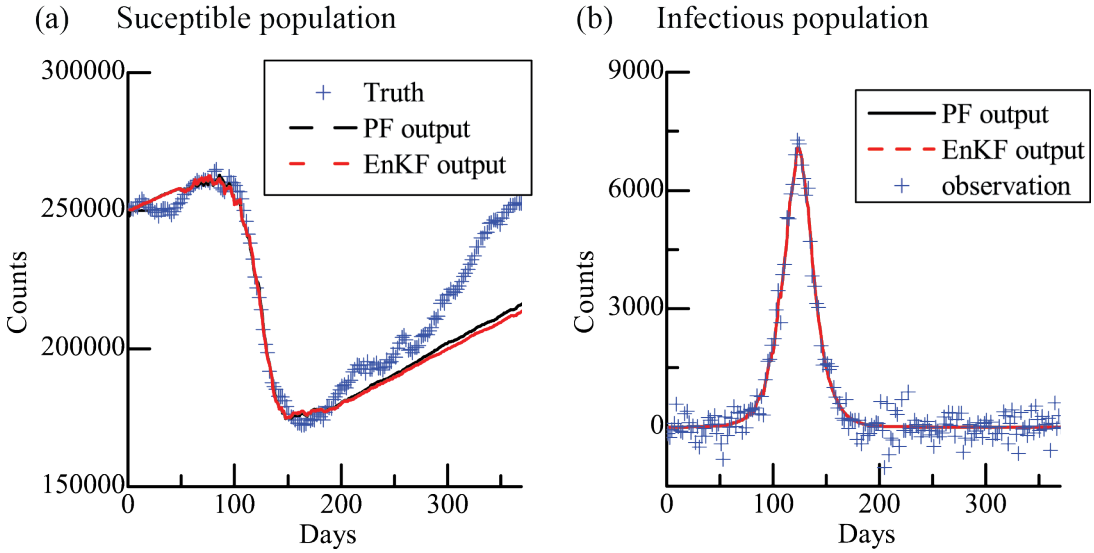

Figure 7: Implementation results of PF and EnKF where the initial estimate for the susceptible population is close to the true value. (a) Susceptible population for the synthetic truth and filtered estimates (b) Infectious population for the synthetic truth and filtered estimates

Figure 8 shows the population estimates using the EKF, UKF and EnKF, where the initial susceptible value is twice the truth. In Figure 8a, the UKF estimate for the susceptible regresses to the true value faster than both the EKF and EnKF estimates. After about 100 days, the estimates from three filters are close to the true value. Figure 8b shows the infectious population estimates, where the UKF now has a most significant error in estimating the infectious population

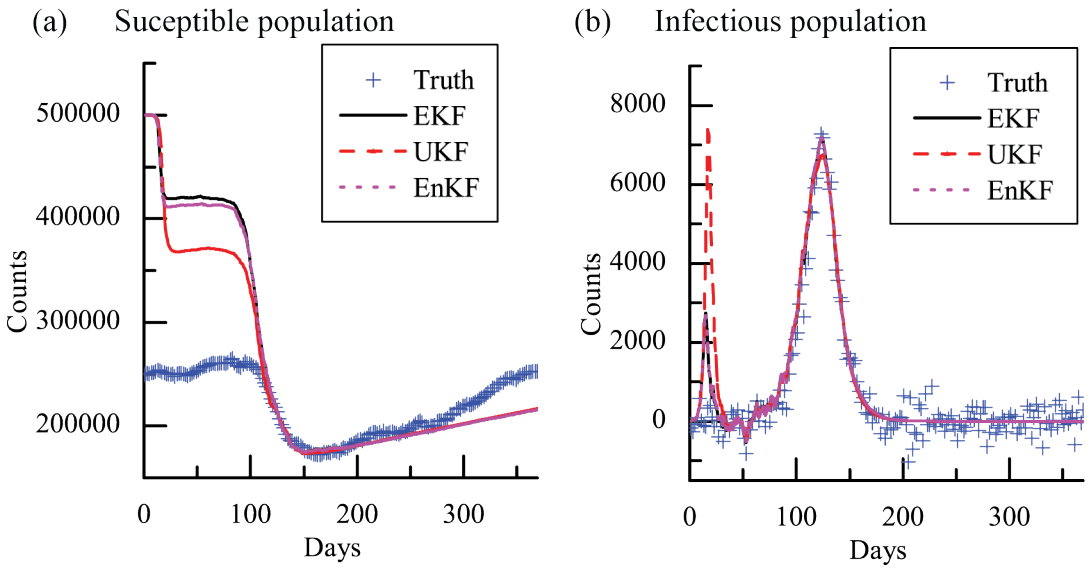

Figure 8: Implementation results of EKF, UKF and EnKF where the initial estimate for the susceptible population is twice the true value. (a) Susceptible population for the synthetic truth and filtered estimates (b) Infectious population for the synthetic truth and filtered estimates. The better estimate of the hidden susceptible population is penalized by a worse estimate of the observed infectious population.

while obtaining the best estimate for the hidden state. A better estimate for the hidden state is penalized by the estimate for the observed state. In the epidemic disease literature, it was reasoned that Kalman filters help to balance the information in the observations and model simulations, when the states (e.g. infectious population, susceptible population) are only partially observable (Shaman & Karspeck, 2012). In contrast, the particle filter method failed to make an estimation of the susceptible population because the importance sampling method (Equation 26) is invalid when the prior distribution is changed. The Kalman filters demonstrate relative robustness even though initial estimate is far from the prior distribution.

To compare the different filter estimates more quantitatively, we run each filter 30 times and obtain the mean square error (MSE) by comparing the filter estimates of the susceptible population with the true values. All these runs start with the prior distribution of the susceptible population, i.e. Gaussian distribution with mean being 250,000 and standard deviation being 1000.

Table 1 lists the mean and variance of the MSE for the 30 runs. UKF has a smaller MSE than EKF, which is not surprising since it has higher order of approximation accuracy than EKF. PF has a slightly smaller MSE than the EnKF. In terms of MSE, UKF is 5% lower than EnKF. The lack of fit in both Figure 6a and Figure 7a, and the large value of MSE are due to the limitation of dynamic system. The susceptible population is not observed and can only be estimated from the relation with the infectious population.

Table 1 also lists the calculation of log likelihood per observation for the different filtering methods. The log likelihood per observation is also calculated for the 30 different runs. In agreement with the MSE comparison, UKF has the highest log likelihood on average even though the variance of log likelihood is not the lowest.

Method	Mean(MSE)	Var(MSE)	Mean(log likelihood per obs.)	Var(log likelihood per obs.)
EKF	1.43E4	1.95E3	-5.11E4	2.02E5
UKF	1.38E4	1.53E3	-1.70E4	2.59E5
PF	1.44E4	1.72E5	-5.16E4	5.33E7
EnKF	1.45E4	2.09E5	-5.27E4	1.86E7

Table 1: Mean square error(MSE) for the filter estimates and the true susceptible population. Mean and Variance for the MSE is obtained by running each experiment 30 times. The log likelihood per observation is also calculated for the 30 different runs.

6.2 Archetype-Filter results for the real influenza data

In the Archetype framework, the process equation (Equation 31) is linear and the observation equation (Equation 33) is not linear. Therefore, the implementation of EKF suffices to make predictions of influenza trend without the complication of sigma points in the UKF method. We implemented both the EKF and PF to forecast the influenza trend of the next week.

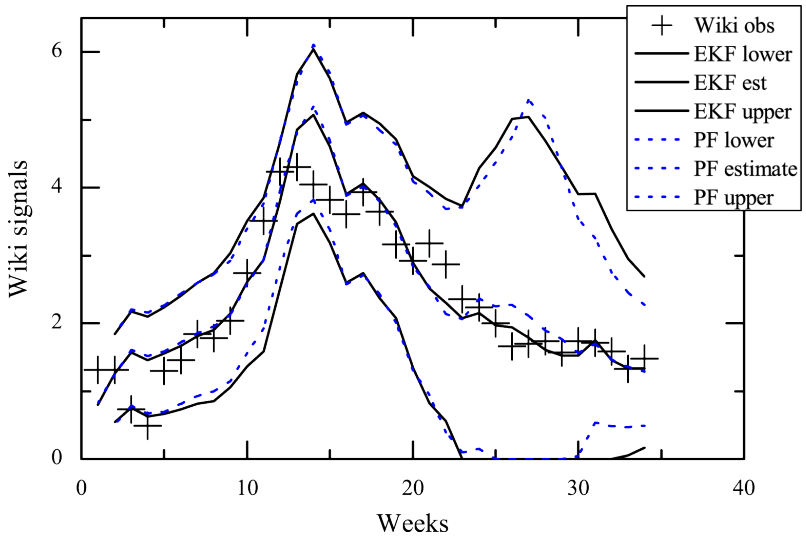

Figure 9: EKF and PF predictions of the Wiki signals of the next week based on all the observation data (Twitter signals, Wiki signals and %wILI) until present time. The curves are the mean, lower and upper bound estimate for one week ahead for the 80% confidence interval. And the + symbol denotes the Wiki signals.

Figure 9 shows the prediction of Wiki signals from EKF and PF for one week ahead. The + symbols represent the observed Wiki signals. Black lines are the EKF prediction of the mean, lower bound and upper bound for the influenza activity in the next week. The dashed blue lines are the PF prediction of the mean, lower bound and upper bound. The bounded regions are of 80% confidence. Both EKF and PF estimates agree with the general trend of the observation values. At the tail of week 25, the EKF curve is closer to the observation value than the PF curve. Furthermore, the peaks from the filter methods are approximately one weeks delayed from the true occurrence time of the influenza peak.

To compare the EKF and PF more quantitatively, we run each filter 30 times and obtain the mean square errors by comparing the predicted Wiki signals with the observed signals.

Method	Mean(MSE)	Var(MSE)
EKF	0.85	0.00044
PF	0.91	0.0050

Table 2: Mean square error(MSE) of the filter prediction in comparison with the observed Wiki signals. Mean and Variance for the MSE is obtained by running each experiment 30 times.

Table 2 lists the mean and variance of the MSE of the 30 runs. We note that EKF has a smaller value of MSE on average as well as a small fluctuation of MSE compared with PF. This quantitative comparison is in agreement with Figure 9.

7 Discussion and Limitation

In the synthetic data, the estimation of the susceptible population has fine agreement with the true values. But if the initial estimate is far away from the truth, it took many iterations to obtain a good estimation of the susceptible population and infectious population as shown in Figure 8. In that sense, SIRS-filter framework is not robust for short term influenza forecast. In reality, to infer the infectious population, the physical parameters including contact rate β, average duration of immunity L, and mean infectious period D also need to be estimated from data. This kind of dual estimation problems for both physical parameters and hidden states requires more careful technical treatments (E. A. Wan et al., 2000). Parameters estimated from historical data may not be useful for the current season as the transmission behavior may be changed. These intrinsic difficulties associated with the SIRS-filter framework may not be avoided in predicting the influenza trend.

The Archetype framework on the other hand makes a simple assumption that future epidemics resemble the past epidemics. To adjust for the uncertainties in influenza peak time and magnitude, two variables (time shift and magnitude scaling) are introduced to forecast the influenza trend. Because of the linearity of the process equation in the Archetype framework, simple implementation of the extended Kalman filter yields good predictions. In comparison with complicated nonlinear process equations, the Archetype framework is not sensitive to the choise of filtering methods. Furthermore, the Archetype framework only requires the two dimensional state vector, thus not computationally expensive to perform the influenza prediction.

From the experimental results, the EKF and PF in the Archetype framework yield good influenza counts for the next week. However, there is a delay of one week for the peak prediction. This delay in forecasting may not be avoided since the prediction is mainly based on the past data of the current year. Without decline in the data, the Archetype framework assume the influenza activity carries some "momentum" into the near future.

8 Conclusions

We implemented four filtering techniques in the experiments of the synthetic data. The influenza trends generated by SIRS mechanistic model were well tracked by the filtering methods. Unscented

Kalman filter yields the best susceptible population estimates in terms of mean square errors and log likelihood, because of the higher approximation accuracy.

In the real influenza case, we predicted the influenza counts for the next week in the Archetype framework. Both the extended Kalman filter and particle filter showed satisfactory performance in forecasting the influenza trend. The 80% bounded region well encloses the real influenza data, with the limitation that the predicted peak is lagged by one week.

9 Acknowledgements

I would like to thank Prof. Roni Rosenfeld, for supervising me during this project. I admire his patience and great explanation power in clarifying ideas and thoughts. I am grateful for Prof. Ryan Tibshirani and Prof. Roy Maxion's comments and questions along the project, which help me to better understand the project. I really appreciate David Farrow on sharing notes and giving a lot of clarifications on his Archetype framework.

References

Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *Signal Processing, IEEE Transactions on*, *50*(2), 174–188.

Brauer, F., Castillo-Chavez, C., & Castillo-Chavez, C. (2001). *Mathematical models in population biology and epidemiology* (Vol. 1). Springer.

Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J., & Rosenfeld, R. (2015, 08). Flexible modeling of epidemics with an empirical bayes framework. *PLoS Comput Biol*, *11*(8), e1004382.

Casti, J. L. (2000). Five more golden rules: knots, codes, chaos, and other great theories of 20th century mathematics. *AMC*, *10*, 12.

CDC. (2012). Overview of influenza surveillance in the united states. *US Centers for Disease Control and Prevention. http://www.cdc.gov/flu/pdf/weekly/overview.pdf*.

Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.-J., & Vespignani, A. (2007). Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS Med*, *4*(1), e13.

Doucet, A., De Freitas, N., & Gordon, N. (2001). An introduction to sequential monte carlo methods. In *Sequential monte carlo methods in practice* (pp. 3–14). Springer.

Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*(12), 656–704.

Dugas, A. F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., & Rothman, R. E. (2013). Influenza forecasting with google flu trends. *PloS one*, *8*(2), e56176.

Evensen, G. (2009). *Data assimilation: the ensemble kalman filter*. Springer Science & Business Media.

Farrow, D. C. (2016). *Modeling the past, present, and future of influenza.* Private Communication, Farrow's work will be part of the published Phd thesis at Carnegie Mellon University.

Gillijns, S., Mendoza, O. B., Chandrasekar, J., Moor, B. L. R. D., Bernstein, D. S., & Ridley, A. (2006). What is the ensemble kalman filter and how well does it work. *American Control Conference*, 6.

Grefenstette, J. J., Brown, S. T., Rosenfeld, R., DePasse, J., Stone, N. T., Cooley, P. C., ... Burke, D. S. (2013). Fred (a framework for reconstructing epidemic dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC public health*, *13*(1), 940.

Grewal, M. S., & Andrews, A. P. (2014). *Kalman filtering: Theory and practice with matlab*. John Wiley & Sons.

Hickmann, K. S., Fairchild, G., Priedhorsky, R., Generous, N., Hyman, J. M., Deshpande, A., & Del Valle, S. Y. (2015). Forecasting the 2013–2014 influenza season using wikipedia. *PLoS Comput Biol*, *11*(5), e1004239.

Ionides, E., Bretó, C., & King, A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, *103*(49), 18438–18443.

Julier, S. J., & Uhlmann, J. K. (1997). New extension of the kalman filter to nonlinear systems. *In AeroSense'97*, *3*, 182–193.

Julier, S. J., & Uhlmann, J. K. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, *92*(3), 401–422.

King, A. A., Ionides, E. L., Pascual, M., & Bouma, M. J. (2008). Inapparent infections and cholera dynamics. *Nature*, *454*(7206), 877–880.

Laporte, R. E. (1993). How to improve monitoring and forecasting of disease patterns. *BMJ*, *307*(6919), 1573–1574.

Mandel, J. (2009). A brief tutorial on the ensemble kalman filter. *arXiv preprint arXiv:0901.3725*.

Maybeck, P. S. (1982). *Stochastic models, estimation, and control* (Vol. 3). Academic press.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Newman, M. E. (2002). Spread of epidemic disease on networks. *Physical review E*, *66*(1), 016128.

Paul, M. J., Dredze, M., & Broniatowski, D. (n.d.). Twitter improves influenza forecasting. *PLoS currents*, *6*.

Quenel, P., & Dab, W. (1998). Influenza A and B epidemic criteria based on time-series analysis of health services surveillance data. *European Journal of Epidemiology*, *14*(3), 275–285.

Shaman, J., & Karspeck, A. (2012). Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, *109*(50), 20425–20430.

Shaman, J., Pitzer, V. E., Viboud, C., Grenfell, B. T., & Lipsitch, M. (2010). Absolute humidity and the seasonal onset of influenza in the continental united states. *PLoS Biol*, *8*(2), e1000316.

Shumway, R. H., & Stoffer, D. S. (2013). *Time series analysis and its applications*. Springer Science & Business Media.

Soebiyanto, R. P., Adimi, F., & Kiang, R. K. (2010). Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. *PloS one*, *5*(3), e9450.

Wan, E., & Van Der Merwe, R. (2000). The unscented kalman filter for nonlinear estimation. In *Adaptive systems for signal processing, communications, and control symposium 2000. as-spcc. the ieee 2000* (pp. 153–158).

Wan, E. A., Van Der Merwe, R., & Nelson, A. T. (2000). Dual estimation and the unscented transformation. *Advances in Neural Information Processing Systems*, 666-672.

Welch, G., & Bishop, G. (2006). An introduction to the kalman filter. 2006. *University of North Carolina: Chapel Hill, North Carolina, US*.

WHO. (2009). Influenza (seasonal). fact sheet no. 211. *World Health Organization, Geneva, Switzerland. http://www.who.int/mediacentre/factsheets/fs211/en/index. html*.

Yang, W., Karspeck, A., & Shaman, J. (2014). Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS Comput Biol*, *10*(4), e1003583.

Yang, W., Lipsitch, M., & Shaman, J. (2015). Inference of seasonal and pandemic influenza transmission dynamics. *Proceedings of the National Academy of Sciences*, *112*(9), 2723–2728.