Kalman Filters in Epidemiological Studies
Directed Reading Program Fall 2020

Lincole Jiang

Mentor: Harrison

Nov. 2020

1 Background Check: What Is an Old-Fashioned, Classic Kalman Filter?

The Kalman filter (KF) was invented by Rudolf E. Kalman, an electrical engineer and mathematician who received the National Metal of Science in 2009. Exclusively derived for linear and Gaussian systems, the classic KF is a two-step process that predicts via a forward model and corrects by updating with available data to estimate the model states of linear systems to derive a best estimate for a linear Gaussian process utilizing only noisy observationsIt is employed under the Bayesian framework wherein unknowns are treated as random variables to sequentially update the posterior distribution of the unknowns conditioned on the observed data [i.e., the mean and the covariance of the underlying Gaussian distribution] using analytically described formulas. It has been widely used in a variety of applications including signal processing, guidance, navigation, and control systems, and perhaps most famously, in the Apollo missions by NASA.

1.1 Brief Description of Classic KF for State Estimation

As stated, the classic KF is a two-step process that both predicts with a forward model and corrects by updating data via an observation model. It all starts with a Gaussian prior distribution:

$$\pi(x_0) \sim N(\bar{x_0}, \Gamma_0)$$

where $\bar{x_0} \in \mathbb{R}^d$ as mean and $\bar{\Gamma_0} \in \mathbb{R}^{d \times d}$ as covariance matrix.

In the prediction step, the state evolution equation that forms the state-space model is assumed to be a linear discrete-time Markov model:

$$X_{j+1} = FX_j + V_{j+1}, V_{j+1} \sim N(0, C)$$

where F is the operation matrix assumed to be constant over time, V_j is a Gaussian random variable that describes the noise process about covariance matrix C, and $j \in 0, 1, ..., j - 1$.

Utilizing this model, we could express the estimated propagation of prior distribution through time via the following analytical equations, which describes the predicted mean and covariance of the underlying Gaussian distribution at $t = j + 1$:

$$X_{j+1|j}^{-} = F x_{j|j}^{-}$$

and
$$\Gamma_{j+1|j}^- = F\Gamma_{j|j}^- + C$$
where $x_{j|j}^-$ and $\Gamma_{j|j}^-$ are the state prediction mean and covariance matrix of the prior distribution, respectively.

Similarly, for the analysis step, the state observation equation is also asumed to be a linear discrete-time Markov model:
$$Y_{j+1} = GX_{j+1} + W_{j+1}, W_{j+1} \sim N(0, D)$$
where W_j is Gaussian random variable with covariance matrix D, G is the operation matrix, and $j \in 0, 1, ..., j - 1$.

With this observation model, we could correct the predicted mean and covariance estimates with the data available via these updating formulas:
$$x_{j+1|j+1}^- = x_{j+1|j}^- + K_{j+1}(y_{j+1} - Gx_{j+1|j}^-)$$
and
$$\Gamma_{j+1|j+1}^- = (I - K_{j+1}G)\Gamma_{j+1_j}^-$$
where $I \in \mathbb{R}^{d \times d}$ is the identity matrix, K_{j+1} is the Kalman gain matrix defined as:
$$K_{j+1} = \Gamma_{j+1|j}G^T(G\Gamma_{j+1|j}G^T + D)^{-1}$$

For each step at time $j+1$, then, completing the two steps leaves the Gaussian posterior distribution, which predicts and updates itself until $t = T$.

2 Ensemble Kalman Filter in Epidemiological Settings

2.1 Why EnKF or Augmented EnKF?

Despite providing the optimal solution in linear systems where assumptions of linearity and Gaussian distributions are met, the classic KF falls short when applying to nonlinear models. By integrating ensemble statistics into the updating equations using an ensemble of discrete realizations from underlying probability distribution to calculate the ensemble mean and covariance, then, a modified filter, called Ensemble Kalman Filter (EnKF), allows for more flexibility for such nonlinear models for estimation of both the states and the observations. Taking a step further, a modified augmented EnKF in the prediction step could track time-varying parameters whereas augmented EnKF is limited to estimating constant parameters. This is called the augmented EnKF for time-varying parameter tracking.

2.2 Brief Description of EnKF

The state space model for EnKF in fact looks very similar to that of the classic KF:
$$X_{j+1} = F(X_j) + V_{j+1}, V_{j+1} \sim N(0, C)$$
$$Y_{j+1} = G(X_{j+1} + W_{j+1}, W_{j+1} \sim N(0, D)$$
where F, G are linear operators. EnKF has similar two-step procedure of predicting and correcting as classic KF, but with discrete samples as probability distribution and each sample propagating

independently. Define S_{j_j} as the discrete sample from the underlying probability distribution at time j s.t. $S_{j|j} = \{x_{j|j}^1, x_{j|j}^2, ..., x_{j|j}^N\}$, with N being the ensemble size and $x_{j|j}^k$ for each $k \in \{1, ..., N\}$. Each individual ensemble member is then updated in each prediction step by state evolution equation:

$$x_{j+1|j}^n = F(x_{j|j}^n) + v_{j+1}^n, n = 1, ..., N$$

with each v_{j+1}^n as realization of the random variable $V_{j+1} \sim N(0, C)$. The analysis step corrects each individual ensemble member via equation:

$$x_{j+1|j+1}^n = x_{j+1|j}^n + K_{j+1}(y_{j+1}^n - \hat{y_{j+1}^n}), n = 1, ..., N$$

where $(y_{j+1}^n = y_{i+1} + w_{j+1}^n$, $w_j + 1 \sim N(0, D)$ is the artificial observation ensemble of size N generated around y_{j+1}; $\hat{y}_{j+1}^n = G(x_{j+1|j}^n)$ the model prediction of the observation, and K_{j+1} the Kalman gain.

2.3 EnKF in Epidemiological Settings

For epidemiological systems, the commonly-employed prediction/forward model is well-established Susceptible-Infectious-Recovered (SIR) model, whose variations had been well-applied from general compartment models for the spread of infectious diseases to more complex, disease-specific models. On the other hand, there are a variety of different observation models or the updating functions available when assessing different levels of complexity. This section presents a brief description of these.

2.3.1 SIR the Prediction model

The SIR model is made up of three compartments describing three different states of a population at a given time t: $S(t)$, the population that are healthy but susceptible to contracting the disease; $I(t)$, the infectious population capable of transmitting the disease to others; and $R(t)$, the population recovered from and has become immune to the disease. If population size N_{pop} is assumed to be constant, the three states of the population could be summarized by formula:

$$R(t) = N_{pop} - S(t) - I(t)$$

via $N_{pop} = R(t) + S(t) + I(t)$.

Whereas the rate of change of susceptible and infectious populations is described by differential equations, respectively:

$$\frac{dS(t)}{dt} = mN_{pop} - \frac{\beta(t)I(t)S(t)}{N_{pop}} - mS(t)$$

$$\frac{dI(t)}{dt} = \frac{\beta(t)I(t)S(t)}{N_{pop}} - \lambda I(t) - mI(t)$$

where m is the constant birth/death rate, $\beta(t)$ is the transmission parameter at time t, and λ is the constant recovery rate. The solution to the stated equations defines the forward propagation in the prediction step of the filter under the framework of augmented EnKF described in the previous section.

2.3.2 Various Observation Functions

Unlike the prediction step, there are a number of available observation functions of varying complexities for different data collection assumptions. This section introduces four of them:

1. Assume that the observed data at time t_i is a direct measurement of $I(t_i)$ at t_i. i.e., $g(x_j, \theta) = I(t_j)$;

2. Assume that the observed data at time t_i is an under-reported measurement of $I(t_i)$. i.e., $(g(x_i, \theta) = \rho I(t_i)$, where ρ is the percentage of population actively reporting cases, assumed constant;

3. Assume that the observed data at time t_i is a measurement of total number of cases accumulated from t_{i-1} to t_i. i.e., $g(x_i, \theta) = \int_{t_{i-1}}^{t_i} \frac{\beta(t)I(t)S(t)}{N_{pop}} \, dt$;

4. Assume that t_i is an under-reported measure of total accumulated cases from t_{i-1} to t_i. i.e., $g(x_i, \theta) = \rho \int_{t_{i-1}}^{t_i} \frac{\beta(t)I(t)S(t)}{N_{pop}} \, dt$.

2.4 Importance of Choice of Observation Model Via Selected Results

When using the true underlying observation function (i.e., 4), the filter is able to estimate the components of the original system with minimal error. However, when the true observation function is not known, using the suboptimal observation functions result in less accurate estimations, with accuracy decreasing as the observation function becomes further from truth (i.e., 1, 2, 3). The study also illustrates that the complexity of the observation function is irrelevant in terms of the accuracy of estimations: similar trends were observed regardless of the complexity of the observation function.

3 References

Mitchell, Leah, and Andrea Arnold. "Analyzing the Effects of Observation Function Selection in Ensemble Kalman Filtering for Epidemic Models." arXiv preprint arXiv:2007.05114 (2020).