From aerospace guidance to COVID-19: Tutorial for the application of the Kalman filter to track COVID-19

Juan I. Langlois * and AnnaMaria B. Dear †
Stanford University, Stanford, CA, 94305

Inspired by the varied responses to COVID-19 among states in the U.S.A, we developed a tutorial on the application of Kalman filtering methods to the estimation of the effective reproduction number. The filter was built considering the SIR epidemiological model and was applied to local infection data for Santa Clara County, California. After estimating the effective reproduction rate from 12 March to 01 November 2020, we analyzed the effectiveness of the public health policies enacted by the Santa Clara Emergency Operations Center in limiting the spread of COVID-19 in the County. Such analysis can be useful to human decision makers who been urged to constrain their policy decisions by the effective reproduction rate of COVID-19 in their areas of jurisdiction.

I. Introduction

While the United States' response to the Coronavirus Disease 2019 (COVID-19) pandemic varied across states in the first half of 2020, the adoption of restrictions on movement and social distancing regulations appeared to help curb the spread of the disease. The aim of this project was to create a tutorial for the application of Kalman filtering to estimate the effective reproduction number, $\mathcal{R}_t$, defined in [1] as "the average number of people infected by a single infectious individual," for the spread of COVID-19. As a case study, the Kalman filter was applied to daily infection data from Santa Clara County.

The standard SIR epidemiological model was used to derive the relationship between the growth rate of infected individuals and the effective reproduction number. Kalman filtering was then applied to smooth the observed growth rate of infected individuals. Because of the linear relationship between the growth rate of infected individuals and the effective reproduction number, Kalman smoothing produced an optimal estimate of the effective reproduction number by minimizing mean-squared error. Infections in a population continue to increase while $\mathcal{R}_t > 1$. Some argue that reaching $\mathcal{R}_t < 1$ should be the constraint placed on public policy during a pandemic [1]. Using this Kalman filtering and smoothing method, human decision makers can analyze the effect of the adoption of policies intended to curb COVID-19 infections on infection growth rates and therefore $\mathcal{R}_t$.

II. Mathematical Model

The estimation of R_t relies on the incorporation of the SIR model to derive the linear relationship between R_t and the growth rate of infected individuals. The standard SIR model in discrete time relates subsets of the total population in a pandemic to one another, describing the evolution of susceptible (S_t), infected (I_t), and recovered (R_t) individuals through the following equations [1]

$$\begin{aligned} S_t &= S_{t-1} - \beta_t I_t \tfrac{S_t - 1}{N} \\ I_t &= I_{t-t} + \beta_t I_{t-1} \tfrac{S_t - 1}{N} - \gamma I_{t-1} \\ R_t &= R_{t-1} + \gamma I_{t-1} \end{aligned} \tag{1}$$

where the total population, N, is equal to the sum of S_t, I_t, and R_t. β_t is the daily disease transmission rate and γ is the daily transition rate from infected to recovered, or recovery rate. Data indicating new infections is used to construct a time series that describes how many individuals are infected at any given time. The transmission rate, β_t is assumed to vary with time because people may make individual choices or be subject to government policies that affect ability to transmit the disease.

*M.S. Candidate, Management Science and Engineering
†M.S. Candidate, Aeronautics and Astronautics

The basic reproduction number, $\mathcal{R}_0^{(t)} \equiv \frac{\beta_t}{\gamma}$ varies over time with β_t. The effective reproduction number is defined as $\mathcal{R}_t = \mathcal{R}_0^{(t)} \frac{S_{t-1}}{N}$. Given the relationships between S_{t-1}, I_t, and β_t in (1), the daily growth rate of the number of infected individuals is

$$gr(I_t) = \frac{(I_t - I_{t-1})}{I_{t-1}}. \tag{2}$$

Using an estimation for the growth rate of infected individuals, denoted by $\hat{gr}(I_t)$, and determining γ, for example through the use of external medical evidence to determine that γ^{-1} is the average infectious period, an estimator for the effective reproduction number can be derived

$$\hat{\mathcal{R}}_t = 1 + \frac{1}{\gamma}\hat{gr}(I_t). \tag{3}$$

The growth rate of I_t can be estimated empirically through the construction of a time series using data on new cases of infection. The time series I_t implied by the relationships in the SIR model is

$$I_t = (1 - \gamma)I_{t-1} + \text{new cases}_t \tag{4}$$

where I_t is initialized with I_0 equal to some constant, the total number of cases of infection on the initial date of the time series. I_t is then constructed recursively. Once the time series is constructed, Kalman filtering tools can be used to smooth the implied $\mathcal{R}_t$ as outlined in the following sections.

III. The Kalman Filter

The Kalman Filter is an example of a recursive Bayesian estimation method that can be used to update a belief distribution over the current unobserved state α_t given the most recent observations y_t. Figure 1 shows a graphical representation of the Bayesian belief network. In the following sections we explain how the Kalman Filter can be used for the filtering and smoothing of the following local level model

$$\begin{aligned} y_t &= \alpha_t + \varepsilon_t \qquad & \text{i.i.d } \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \\ \alpha_{t+1} &= \alpha_t + \eta_t \qquad & \text{i.i.d } \eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \end{aligned} \tag{5}$$

To apply this model to our COVID problem we can use $y_t = f(gr(I_t)) = 1 + gr(I_t)/\gamma$ and $\alpha_t = \mathcal{R}_t$. The rest of this section is a summary of the main expressions and concepts necessary to implement a Kalman filter for the local model as presented in [2].

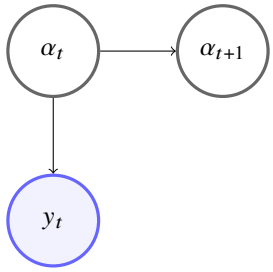

Fig. 1 Local-level model Bayesian belief network

A. State filtering

The objective of filtering is to update our knowledge of the state variables each time a new observation y_t is brought in. Let Y_{t-1} be the vector of observations $(y_1, \cdots, y_{t-1})^T$ for $t = 2, 3, \cdots$ and assume that the conditional distribution of α_t given Y_{t-1} is $\mathcal{N}(a_t, P_t)$ where a_t and P_t are known. Assume also that the conditional distribution of α_t given Y_t is $\mathcal{N}(a_{t|t}, P_{t|t})$. The distribution of α_{t+1} given Y_t is $\mathcal{N}(a_{t+1}, P_{t+1})$.

The objective of the filtering process is to calculate $a_{t|t}$, $P_{t|t}$, a_{t+1} and P_{t+1} when y_{t+1} is observed. Generally $a_{t|t}$ is referred as the *filtered estimator* of the state α_t and a_{t+1} as the *one-step ahead predictor* of α_{t+1}. Their respective associated variances are $P_{t|t}$ and P_{t+1}. Finally, the last important definition is the one-step ahead prediction error v_t of y_t with its corresponding variance F_t. The prediction error is also known as an *innovation*.

With all these definitions in mind, the Kalman Filter for the local level model can be summarized by the following recursive relations

$$\begin{aligned} v_t &= y_t - a_t & F_t &= P_t + \sigma_\varepsilon^2 \\ a_{t|t} &= a_t + K_t v_t & P_{t|t} &= P_t(1 - K_t) \\ a_{t+1} &= a_t + K_t v_t & P_{t+1} &= P_t(1 - K_t) + \sigma_\eta^2 \end{aligned} \tag{6}$$

for $t = 1, \cdots, n$, where $K_t = P_t / F_t$ is the known as the *Kalman gain*. It is interesting to observe how the Kalman filter reconciles the observed information and the model predictions to update our beliefs. For this simple model it is easy to see that the Kalman gain is a type of signal-to-noise ratio and is used to determine how much weight to give to the new observation. If we have very noisy observation process we assign less weight to the observation and more weight to the prediction given by the dynamic model in the previous iteration.

B. State smoothing

In the previous section we estimated the latent states given the past data and the most recent observation. Now we turn present the operation of *state smoothing*, where we estimate the values $\alpha_1, \cdots, \alpha_n$ using all the data Y_t. The resulting estimates $\hat{\alpha}$ and V_t are known as the is the *smoothed state* and *smoothed state variance*, respectively.

In order to estimate the smoothed state and variance we need to define a new gain $L_t = 1 - K_t$, which is used for the weighted sum of innovations or *smoothing cumulant* r_t and weighted sum of the inverse variances of innovations or *smoothing variance cumulant* N_t backward recursions

$$r_{t-1} = \frac{v_t}{F_t} + L_t r_t \qquad N_{t-1} = \frac{1}{F_t} + L_t^2 N_t \tag{7}$$

Since no values are available after time n, we can set $r_n = 0$ and $N_n = 0$. The computation of r_t and N_t for $t = 1, \cdots, n-1$ are straight forward following the backward recursions.

Finally, the smoothed state and variance are given by the following relations

$$\hat{\alpha}_t = a_t + P_t r_{t-1} \qquad V_t = P_t - P_t^2 N_{t-1} \tag{8}$$

IV. Initialization

Thus far we have assumed that the distribution of the initial state α_1 is $\mathcal{N}(a_1, P_1)$ where a_1 and P_1 are known. However, it is the general case that we do not know anything about the distribution of α_1. In this situation it is reasonable to represent α_1 as having a *diffuse prior* density, that is, fix a_1 at an arbitrary value and let $P_1 \to \infty$.

This process is known as a *diffuse initialisation* of the Kalman filter and the resulting filter is called *the diffuse Kalman filter*. This initialization yields the same values of a_t and P_t for $t = 2, \cdots, n$ that would be obtained by treating y_1 as fixed and taking $\alpha_1 \sim \mathcal{N}(y_1, \sigma_\varepsilon^2)$, which is intuitively reasonable in the absence of information about the marginal distribution of α_1.

V. Parameter estimation

For the previous sections we have assumed that the parameters σ_ε^2 and σ_η^2 are known. This is not the case in practice, and these parameters have to be fitted from real data. This can be done using classical inference by means of maximum likelihood. In this case the maximum likelihood looks like

$$L = p(y_1, \cdots, y_t)$$

Given the structure of our belief network, the joint density can be expressed as

$$P(Y_t) = \prod_{t=1}^{n} p(y_t|Y_{t-1})$$

where $p(y_1|Y_0) = p(y_1)$. Now $p(y_t|Y_{t-1}) = \mathcal{N}(a_t, F_t)$ and $v_t = y_t - a_t$, so on taking logs and assuming that a_1 and P_1 are known the loglikelihood is given by

$$\log L = \log p(Y_n) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\left(\log F_t + \frac{v_t^2}{F_t}\right) \tag{9}$$

The exact loglikelihood can therefore be evaluated easily from the Kalman filter. The unknown parameters σ_ε^2 and σ_η^2 can be estimated by maximizing expression 9 numerically. In each iteration k of the numerical maximization we have to use the new estimates of $\hat{\sigma}_\varepsilon^{2(k)}$ and $\sigma_\eta^{2(k)}$ to run the recursions of expression 6 and evaluate a new value of the loglikelihood $\log L^{(k)}$.

VI. Application to real data

We now apply the Kalman Filter to real data from the Santa Clara County in California. The dataset was downloaded from the official county website*. In order to have a significant starting point, we considered the data after the total number of infected individuals was greater than 100. The number of active infections I_t was then determined for $t = 1, \cdots, n$ through the following recursion from (4)

$$I_t = (1 - \gamma)I_{t-1} + \text{new cases}_t$$

with $I_0 = 100$. The growth rate of infections was then estimated according to the expression 2. The resulting time series had observations from March 12th, 2020 to November 1st, 2020. Figure 2 shows the state variables of active infected cases I_t and the associated growth rate $gr(I_t)$. The negative values in the growth rate can be explained by the fact that the number of daily new cases observed is less than the number of daily recovered given by γI_t. From looking at the data it is clear that the infected growth rate is an extremely noisy measurement and decision makers could benefit greatly by the use of filtering and smoothing methods.

The maximum likelihood estimates for the error and random walk terms with a diffuse initialisation where $\hat{\sigma}_\varepsilon^2 = 0.2193$ and $\hat{\sigma}_\eta^2 = 0.0040$. This implies that the effective reproduction rate has a very low signal-to-noise ratio ($\hat{\sigma}\varepsilon^2/\hat{\sigma}\eta^2$) of 0.0184. This estimate is consistent with the observations of Figure 2. The computed filtered and smoothed state are plotted in figures 3 and 4, respectively.

VII. Discussion

For the discussion, we would like to do a retrospective evaluation of the different policy actions that were taken by the policymakers of Santa Clara County and how they impacted the effective reproduction number. However, we first would like to comment on the quality of the filtered and smoothed state estimates. The filtered values, although significantly more stable than direct observations, are still very volatile. The smoothed output is a more robust estimation of the underlying states and has tighter confidence bounds. This is expected since the latter output uses all the observation, past and future, for the state estimates, whereas the filtering process only uses past information.

The main differences in state estimates are for the middle of the period. For the beginning and end of the time series, both estimates yield similar results. This implies that for an online planning strategy the smoothed state estimates might not contribute additional information. For a retrospective analysis, the smoothed state estimates are a better reference for the true unobserved state values.

*https://data.sccgov.org/browse?category=COVID-19

In order to asses the policy executed by local authorities we reviewed the dates and implications of different health orders issued by the County of Santa Clara Emergency Operation Center[†]. Specifically, we focused on orders that restricted the activities, businesses and general mobility of the County residents. The first Order to Shelter in Place was issued 03-26-2020 and was the most strict. After this order, the Emergency Operation Center issued two additional health orders, on 04-29-2020 and 06-05-2020, that revised and extended the original shelter in place order. The revisions and extensions determined the new conditions for the shelter in place order. In general, subsequent revisions were more lenient. It should be noted that the 06-05-2020 order which allowed for the re-opening of many non-essential businesses, including outdoor dining at restaurants, indoor shopping, in-home services, and childcare and summer camps, eased restrictions before $\mathcal{R}_t < 1$ and preceded a significant increase in cases of COVID-19 which peaked in early July. Finally, the last important order was issued on 07-02-2020, which established risk reduction measures implied by a shift in approach.

In order to account for the residents' actual behavior, we included the mobility index used in [1]. This mobility index was constructed from the "COVID Community Mobility Reports of Google." The mobility reports included changes in mobility for six different sectors for the Santa Clara County. These changes could be aggregated into one significant mobility index by using Principal Component Analysis. The mobility index obtained with this method accounts for more than 60% of the mobility dataset variance.

Figure 5 shows this estimated mobility index plotted against the smoothed effective reproduction number. From the time series it is clear that the shelter in place policy was extremely effective in lowering the spread of COVID. The mobility index shows that this action was effective in reducing the number of in person interactions between residents after this period. We can also observe that after the shelter in place revisions, the residents mobility quickly increased and with it the transmission of the virus. Finally, after the health order establishing the new risk reduction measures, the county was able to reduce the effectively reproduction number to levels of around $\mathcal{R} = 1$ with out further restricting the residents mobility.

VIII. Conclusion

The objective of this project was to develop a tutorial for the application of the Kalman filter to a COVID-19 time series. The applied methodology allows decision makers to arrive at reliable estimates of the effective reproduction number of the disease, a dynamic parameter that is directly related to the infection's transmission rate and is therefore a useful control variable. As part of this tutorial we applied Kalman filtering and smoothing to daily data from Santa Clara county, and analyzed how different health policy actions affected the effective reproduction number. The results clearly show that Kalman filtering is an important tool that can be used by decision makers dealing with uncertain phenomenons.

References

[1] F., A.-M., F., B., Kucunskas, and Rondon-Moreno, "Tracking R of COVID-19: A New Real-Time Estimation Using the Kalman Filter," *Available at SSNR*, 2020. https://doi.org/10.2139/ssrn.3581633.

[2] J., D., and J., K. S., *Time Series Analysis by State Saace Methods*, 2nd ed., Oxford University Press, Oxford, 2012, Chap. 2.

[†]https://www.sccgov.org/sites/covid19/Pages/public-health-orders.aspx#executive-summary

Appendix - Plots

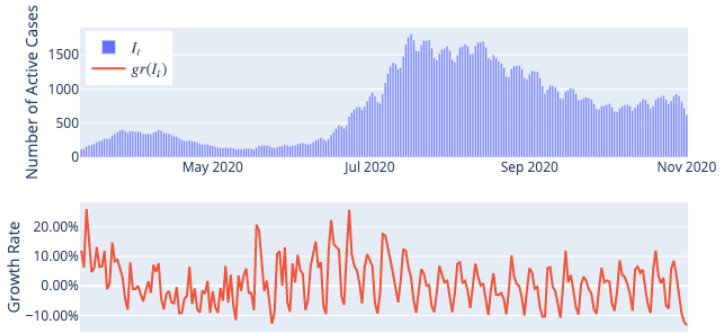

Fig. 2 Total number of active cases and associated growth rate as calculated from the reported daily new cases for Santa Clara county. These state variables were constructed considering an average infectious period of $\gamma^{-1} = 7$.

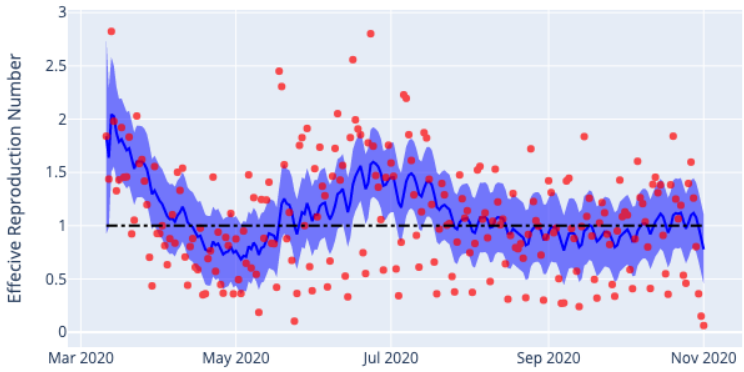

Fig. 3 Implied observed $\mathcal{R}_t$ (red markers) and the output of state filtering recursion (blue solid line). We also included 95% the filtered state confidence interval and a reference line for $\mathcal{R} = 1$ (black dashed line).

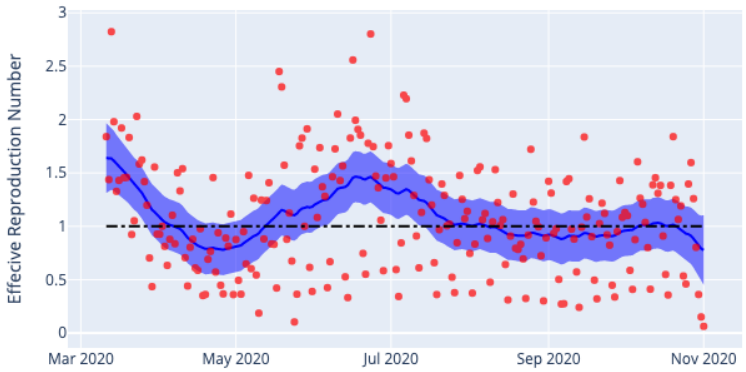

Fig. 4 Implied observed $\mathcal{R}_t$ (red markers) and the output of state smoothing recursion (blue solid line). We also included the 95% smoothed state confidence interval and a reference line for $\mathcal{R} = 1$ (black dashed line).

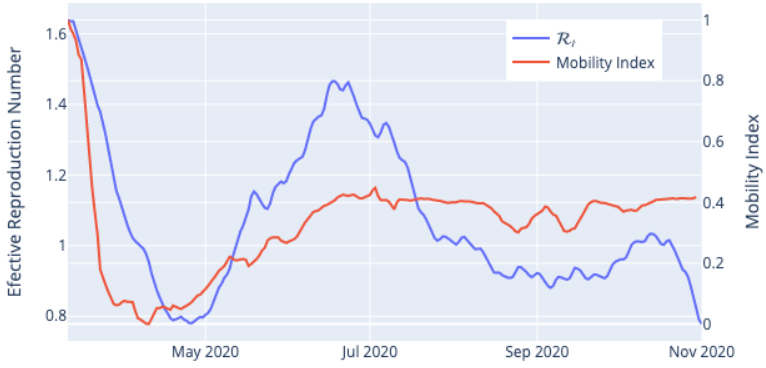

Fig. 5 Time series of the smoothed effective reproduction number and the mobility index. For interpretability, we normalized the mobility index to a unit range. we also applied a 7 day moving average to smooth spikes do to weekend activity.