Volatility estimation for COVID-19 daily rates using Kalman filtering technique

Md Al Masum Bhuiyan [a], Suhail Mahmud [b,*], Md Romyull Islam [c], Nishat Tasnim [c]

[a] *Austin Peay State University, USA*
[b] *The Pennsylvania State University, USA*
[c] *Daffodil International University, Bangladesh*

A B S T R A C T

This paper discusses the use of stochastic modeling in the prognosis of Corona Virus-Infected Disease 2019 (COVID-19) cases. COVID-19 is a new disease that is highly infectious and dangerous. It has deeply shaken the world, claiming the lives of over a million people and bringing the world to a lockdown. So, the early detection of COVID is essential for the patients' timely treatment and preventive measures. A filtering technique with time-varying parameters is presented to predict the stochastic volatility (SV) of COVID-19 cases. The time-varying parameters are estimated using the Kalman filtering technique based on the stochastic component of data volatility. Kalman filtering is essential as it removes insignificant information from the data. We forecast one-step-ahead predicted volatility with ± 3 standard prediction errors, which is implemented by Maximum Likelihood Estimation. We conclude that Kalman filtering in conjunction with the SV model is a reliable predictive model for COVID-19 since it is less constrained by the past autoregressive information.

Introduction

Forecasting of time series with the estimation of time-varying parameters is useful for many statistical, probabilistic, and optimization processes that allow models to consider past observations and detect the disease pattern. Researchers and developers are increasingly using stochastic models to track and prevent chronological diseases and gain a more comprehensive understanding of the disease. Recently, many researchers, journalists, and amateur data enthusiasts are working on stochastic models to help people monitor the Coronavirus's spread and effects over time.

Corona Virus is a respiratory illness caused by a novel virus that affects humans, mammals, and birds. This viral disease has become a major global disaster. Novel Coronavirus outbreaks were initially detected in Wuhan, China (in 2019), and have now spread to numerous countries worldwide. Nearly 40 million cases have been identified in 188 countries by October 15, 2020, with over one million fatalities and 27 million recovering [1].

Experts have confirmed that the virus can spread rapidly from one human body to another and infect the lungs of humans through the respiratory system. Close to each other (less than six feet), the virus spreads through droplets generated from coughing, sneezing, and talking. Most of the droplets fall to the ground or onto surfaces rather than traveling long distances in the air. Individuals who have been infected with this viral infection will have varying symptoms, from coughing, fever, infections in the throat, kidney failure, respiratory problems, etc. The less common way in which people become infected is by touching something contaminated then touching their faces. It is most contagious during the first three days after the onset of symptoms, even though people who do not show symptoms are also at risk of picking it up before symptoms appear aka being asymptomatic [2,3].

COVID-19 has been affecting the US for months, and researchers are working hard to determine the virus's characteristics (why some people are more affected than others, what we can do to slow its spread, and where it is likely to move to next). The data indicate a spike within a short time, therefore it might be useful to analyze the disease case rate to know how the disease is spreading, what impact the pandemic has on people, and whether the preventative measures are effective [4,5]. The dynamics of the COVID-19 cases are now believed to involve volatility clustering and show typical non-linear characteristics [22].

This study develops a stochastic model to predict the volatility of COVID-19 rates per day. The volatility models are used to predict the financial data since they contain extreme fluctuations. The SV model with Kalman filtering is used in this analysis because COVID-19 data

shows high spikes within a short time period. A challenge of the SV models for COVID-19 data is the estimation of time-varying parameters, because it is not possible to observe the volatility directly from the data. Our approach consists of observing only a time series of daily rates, followed by a filtering procedure. At this point, the rates are not a Markov process, the likelihood of a current observation of rates is a function of the entire history of the COVID cases, not just the last observation. However, the rates of infected cases follow long memory nature over time. It suggests that there is a persistent behavior from the past to present information. We assumed a log-volatility (conditional) followed by an auto-regressive stationary process in the COVID-19 data. The long memory and stationarity are verified by testing the parameters and the unit root test presented later. As the likelihood estimation of the SV model is tiresome for fitting the data, the Kalman filtering has been used to estimate the time-varying parameters via Maximum Likelihood Estimation [9].

The overview of this paper is as follows: Section 2 describes the research methodology of Kalman filtering [16] and volatility models. Section 3 deals with the dynamic behavior of the datasets. We discuss the background and some useful information about COVID-19. In Section 4, the long memory and stationarity tests are analyzed. Section 5 provides the results and discussion of our model's suitability regarding the estimation of model parameters for COVID-19 data. Finally, Section 6 contains the conclusion of this study.

Research methodology

This section describes the Kalman Filtering technique and volatility modeling to estimate the stochastic volatility (SV) of COVID cases daily rates.

Kalman filtering

We begin with a state-space model [14] as

$$z_t = C\theta_t + \epsilon_t \tag{1}$$

where z_t is observed data (space), θ_t is unobserved data (state vector) with a coefficient matrix C, and ϵ_t is a Gaussian error term. As θ_t is unobserved, we use the following autoregressive equation:

$$\theta_t = B\theta_{t-1} + v_t, \tag{2}$$

where B is a $n \times n$ transition matrix and v_t is a Gaussian error term with mean 0 and variance σ_v. The unobserved data θ_t can be obtained from Eqs. (1) and (2) using given data $z_s = \{z_1, ..., z_s\}$. In this study, time s is used as time t in a recursive process and the process for t is called as filtering technique. The filtering technique helps to find accurate estimation from noisy information. The error terms η_t can be defined as $\eta_t = \widehat{\theta}_t - \theta_t$ using the unobserved data. Besides the error terms, the covariances of the two noise terms are assumed as stationary over time as $E[v_t v_t^T] = \sigma_v$ and $E[\epsilon_t \epsilon_t^T] = \sigma_\epsilon$. The best filter can be found by minimizing the mean squared error, $E(\eta_t^2)$, which is equivalent to W_t (the error covariance matrix at time t):

$$W_t = E[\eta_t \eta_t^T] = E[(\theta_t - \widehat{\theta}_t)(\theta_t - \widehat{\theta}_t)^T] \tag{3}$$

Now the state vector can be updated with an innovation process. The $\widehat{\theta}_t'$ is assumed as the prior estimate of $\widehat{\theta}_t$, which is computed by the state update equation as

$$\widehat{\theta}_t = \widehat{\theta}_t' + K_t(z_t - C\widehat{\theta}_t') \tag{4}$$

where $\widehat{\theta}_t'$ is the prior estimate of $\widehat{\theta}_t$, K_t is the Kalman gain and $(z_t - C\widehat{\theta}_t')$ is the innovation or measurement residual. At this point, the error covariance matrix at time t can be obtained as

$$W_t = W_t' - K_t C W_t' - W_t' C^T K_t^T + K_t(C W_t' C^T + \sigma_\epsilon)K_t^T \tag{5}$$

It is clear that the error of the prior estimate is not correlated with the innovation. To minimize the tr($[W_t]$), we differentiate it with respect to K_t and set it to zero as

$$\frac{dtr[W_t]}{dK_t} = -2(CW_t')^T + 2K_t(CW_t'C^T + \sigma_\epsilon) = 0 \tag{6}$$

In order to compute Kalman gain K_t, we minimize the trace of W_t by taking derivative as equal 0, which gives:

$$K_t = W_t' C^T (CW_t' C^T + \sigma_\epsilon)^{-1} \tag{7}$$

Eq. (7) is called as Kalman gain equation in the filtering technique. Now Eq. (5) can be updated with the optimal gain as

$$W_t = W_t' - K_t C W_t' = (I - K_t C)W_t' \tag{8}$$

The prior estimate of W_t can be expressed as

$$W_{t+1}' = \phi W_t \phi^T + \sigma_v \tag{9}$$

where ϕ is the stationary transition matrix and the above Eq. (9) gives the minimum MSE. For the details of filtering, the readers are referred to [12].

Volatility modeling

This subsection presents the volatility modeling of daily rates of COVID-19 data. A stochastic component has been used to compute the volatility by following an innovation sequence. The innovation is fully independent of observations used in this study [13]. The data volatility is estimated through an unobservable process that changes stochastically. We express the rates r_t as a product of two components of the process as

$$r_t = \sigma_t \beta_t, \tag{10}$$

where σ_t is the volatility and β_t is a noise term. At this point, we assume that the noise term follows a sequence of Gaussian white noise [15], and there is no dependency between the sequence and data volatility.

To estimate the stochastic volatility, the log-squared rates of the data are used as follows:

$$\psi_t = v_t + \alpha_t, \tag{11}$$

where $\psi_t = \log r_t^2$, $v_t = \log \sigma_t^2$, and $\alpha_t = \log \beta_t^2$. Now it is clear that the log-squared rates have two parts namely, the unobserved volatility v_t and the unobserved noise α_t. The unobserved volatility varies with time through an auto-regression equation [11]:

$$v_t = a_0 + a_1 v_{t-1} + \gamma_t, \tag{12}$$

where γ_t is a white Gaussian noise term with the variance σ_γ^2. So the Eqs. (11) and (12) consist of the time-varying parameters and are called the stochastic volatility model. In this model, the noise term is computed by two types of Normal distribution. One of the Normal distributions is assumed with zero mean and the other one is non-zero mean. So the observed data z_t can be expressed as follows:

$$z_t = \lambda + v_t + \delta_t, \tag{13}$$

The noise part δ_t can be written as a linear combination of Bernoulli random variable [20], with probability π and Normal random variables x_{t0} and x_{t1}, where x_{t0} is Normally distributed with mean 0 and variance u_0^2, x_{t1} is Normally distributed with mean μ_1 and variance u_1^2 with $\delta_t = B_t x_{t0} - (B_t - 1)x_{t1}$. In our study, we assume that x_{t0}, x_{t1}, B_t all are independently and identically distributed and the probabilities of B_t are
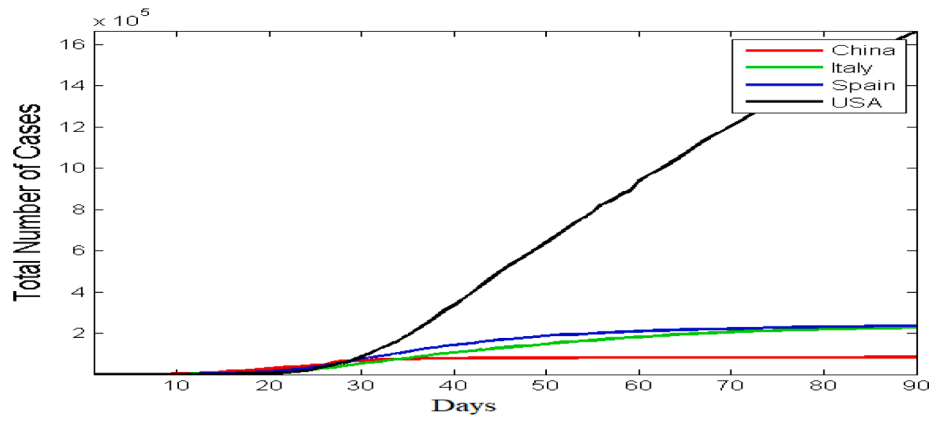
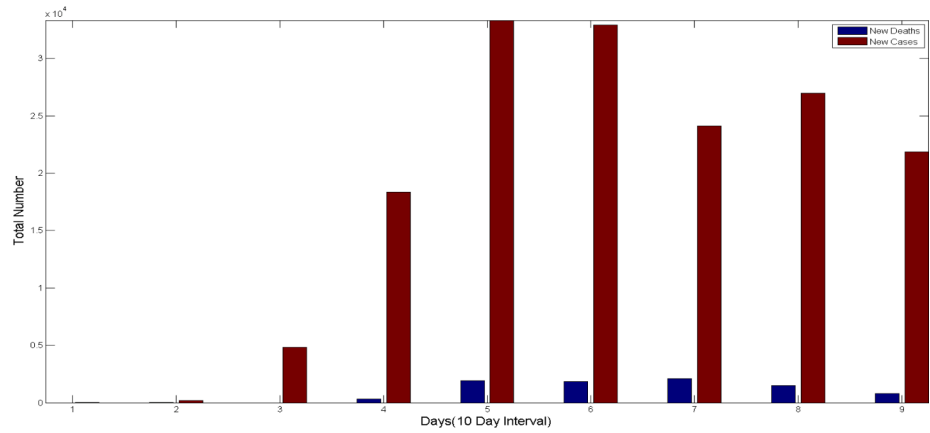**Fig. 1.** Comparison of countries for daily cases.

Fig. 2. Daily new deaths and new cases for 10-day interval time interval in USA for first 90 days.
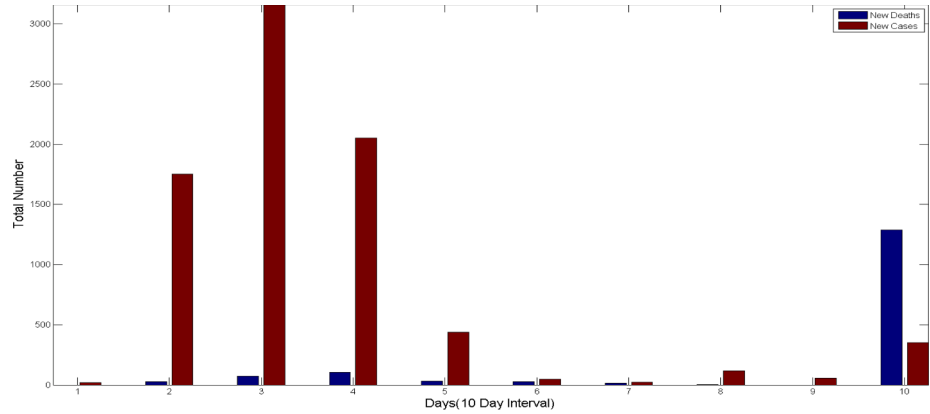

Fig. 3. Daily new deaths and new cases for 10-day time interval in China.

defined as $\Pr\{B_t = 0\} = \pi_0$ and $\Pr\{B_t = 1\} = \pi_1$, where $\pi_0 + \pi_1 = 1$. In the SV model, $a_0, a_1, \sigma_\gamma, \mu_1, u_0$, and u_1 are time-varying parameters, so our approach is to estimate them using the Kalman filtering technique described in Section "Kalman filtering".

Dynamic behavior of the datasets

In this section, we present the background of the COVID-19 datasets used in the paper. It is the dynamic behavior of the data that encourages us to apply our methodology in this paper.

Data background

We collected the daily number of laboratory and hospital confirmed COVID-19 cases and deaths released by the World Health Organization (WHO) from January 10, 2020 to May 15, 2020 to construct a real-time database [6]. Most affected countries like the United States, China, Italy, and Spain were included in this study and a comparison of daily cases was illustrated (See Fig. 1). Afterward, daily new deaths and new cases for a 10-day interval time interval in all four countries were plotted for the first 90 days (See Figs. 2–5). Healthcare Access and Quality (HAQ) Index for the topmost affected countries with confirmed COVID-19 cases
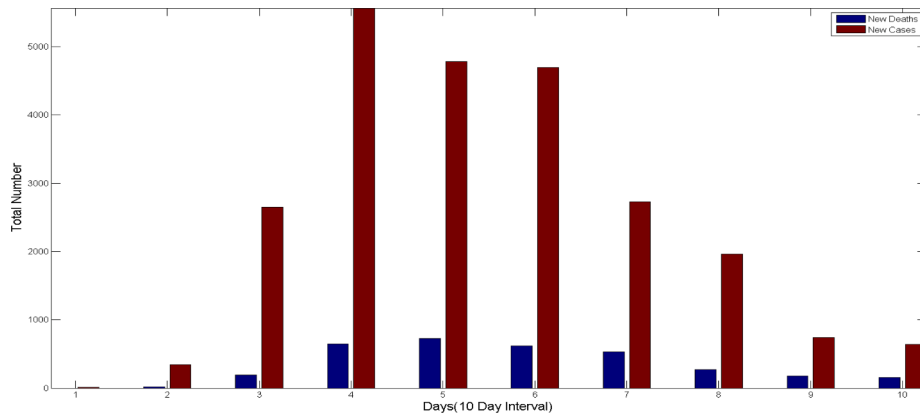
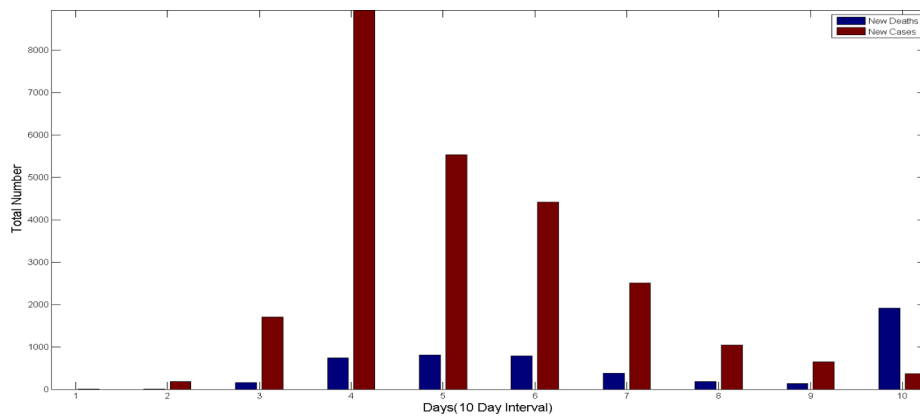**Fig. 4.** Daily new deaths and new cases for 10-day time interval in Itlay.

Fig. 5. Daily new deaths and new cases for 10-day time interval in Spain.

Table 1
Descriptive statistics of COVID-19 dataset.

Statistics	USA	China	Italy	Spain
Mean	0.2291	0.2844	0.1051	0.0353
Std. dev	0.7511	1.3633	0.5108	0.8016
Minimum	−0.5822	−1	−0.4563	-4.1108
Maximum	5.6671	7	3.4285	3.5975
Skewness	5.8577	3.8789	3.9437	−0.8655
Kurtosis	41.8821	16.3303	20.7222	12.7746

Table 2
Unit root testing (ADF) for COVID cases.

Country	Test statistics	Lag	p-value
USA	−2.96	4	0.1779
Spain	−3.03	4	0.1491
Italy	−3.51	4	0.0441
China	−3.10	4	0.1187

Stationary and long memory approaches

This section analyzes the time series by testing for stationarity and long memory in the COVID-19 cases data. A stationary series and long memory series are relatively easy to predict. The assumption is that the data's statistical properties will be the same in the future as they were in the past. We now briefly discuss the long memory and stationary test when they are applied to the datasets.

Stationary test

To test COVID-19 data's stationarity, we used the Augmented Dicky Fuller test (ADF test) [21]. It is a hypothesis used to determine the presence of a unit root in a series which facilitates the analysis of higher-order autoregressive processes. The null hypothesis is assumed as the data has a unit root against the alternative with no unit root. The p-value below the critical level leads to a unit root in the dataset [17]. This test's summary statistics for the datasets are presented in Table 2.

We see that all the p-values of four datasets are higher than a significance level (0.01) at lag 4, which suggests that the alternative hypothesis is acceptable with no unit root, meaning that the datasets of

reported by the WHO is derived from a previously published study by the GBD (Global Burden of Disease) 2016 Healthcare Access and Quality Collaborators [7].

Descriptive statistics

This data set includes data from four different countries: the United States, China, Italy, and Spain. We assemble daily new cases and new deaths for the first 90 days for each country and calculate the change in percentage [8]. Next, we applied some statistical analysis to the datasets for additional information. Table 1 provides information about the percentage change of daily new cases for each country. Spain has a lower mean value than other countries. The standard deviation for the maximum cases is higher in China, as presented in Table 1. The skewness and kurtosis give summary information about the shape of a distribution. As it shows that the Kurtosis is positive for all the countries and maximum for the USA, it indicates flatter tails and narrow peaks aka normal distribution.

Table 3
Estimates of long-memory parameter for COVID cases data.

Source	Estimates (long memory parameter)	Std. Error
USA	0.388	0.081
Spain	0.457	0.083
Italy	0.457	0.083
China	0.266	0.069

Table 4
Estimates of SV parameter for the COVID cases in USA.

Model parameters	Estimate	Standard Error
a_0	0.0017	0.048
a_1	0.991	0.032
σ_γ	0.399	0.037
u_0	−0.0064	0.082
μ_1	−6.848	1.7400
u_1	6.432	1.1641

Table 5
Estimates of SV parameter for the COVID cases in Spain.

Model parameters	Estimate	Standard Error
a_0	0.223	0.198
a_1	0.986	0.020
σ_γ	0.742	0.114
u_0	−0.0026	0.140
μ_1	2.573	1.234
u_1	4.520	0.880

COVID-19 are stationary at lag 4.

Long memory test

As the data follow stationary behavior at a specific lag, it prompts us to analyze the long memory effects of data. We know that the fractional difference parameter identifies a long memory pattern of data in an Autoregressive Fractional Integrated Moving Average (ARFIMA) model [18]. The process is considered the long memory pattern when the fractional difference parameter (known as long memory parameter) lies in the interval (0, 0.5). Since the parametric model ARFIMA was fitted to the Gaussian stationary data r_t, a traditional Maximum Likelihood (ML) can be used to estimate the model parameter. However, we observe that the traditional ML estimator (MLE) requires a large number of operations to optimize its likelihood function for a Gaussian random field, thus it is not computationally efficient. At this point, we used a relatively efficient algorithm namely, the whittle likelihood that provides a spectral approximation to the log-likelihood [19]. Using the stationary property, the whittle approximation can reduce the MLE's number of operations from $\mathcal{O}(n^3)$ to $\mathcal{O}(nlogn)$. Table 3 shows the parameter estimation with standard error for COVID data from four countries. We see that the estimated parameter is less than 0.5 for each country's dataset. Furthermore, the estimated errors are very low, meaning that the estimates are stable around the actual value. So the datasets used in the study follow long memory patterns, i.e., persistence behavior.

Results & discussion

This section presents the analysis of estimating the time-varying parameters of SV model using the Kalman filtering technique. To fit the Kalman filtering, we first initialize the parameters $a_0, a_1, \sigma_\gamma, \mu_1, u_0$, and u_1 for estimation. The initialization was considered in a way to obtain the log-volatility over time. The parameter σ_γ^2 represents the

Table 6
Estimates of SV parameter for the COVID cases in Italy.

Model parameter	Estimate	Standard Error
a_0	0.238	0.165
a_1	0.988	0.018
σ_γ	0.553	0.069
u_0	0.0036	0.0878
μ_1	3.637	1.302
u_1	5.112	1.007

Table 7
Estimates of SV parameter for the COVID cases in China.

Model parameter	Estimate	Standard Error
a_0	−0.271	0.199
a_1	0.951	0.036
σ_γ	−0.403	0.074
u_0	0.0001	0.954
μ_1	1.656	1.284
u_1	4.920	0.1171

variance of the log-volatility process and measures the randomness of future data volatility. To estimate the parameters at time t, the MLE algorithm was used with the innovation processes in Eqs. (12) and (13). In this case, we used the normally distributed auto-regressive conditional heteroscedasticity assumption on the white noise term, β_t [10].

The parameter estimates ($a_0, a_1, \sigma_\gamma, \mu_1, u_0$, and u_1) and the sample paths of data volatility after filtering are presented in Tables 4–7 and Figs. 6–9. It is clear that the estimates are close to the true parameters, as the errors are pretty low. The σ_γ^2 is the variance parameter of the log-volatility process, which measures the uncertainty of future data volatility. If the value of σ_γ^2 is zero, it is not possible to identify the SV model. The parameter a_1 is considered as a measure of the persistence of shocks to the volatility. Tables 4–7 show that a_1 is less than 1 for four countries' data volatility. So we conclude that the latent volatility process is stationary, leading to the stationarity of z_t of the COVID cases, which confirms the results of Section "Stationary and long memory approaches".

The parameters a_1 and σ_γ represent the dynamics of the volatility evolution of COVID cases. The tables also show that the parameter a_1 is pretty close to 1, and the parameter σ_γ is different from 0 for all four countries. It suggests that the volatility evolution is uneven over time. It is concluded that the COVID rates of cases might be heteroscedastic by nature, meaning that there might be non-constant conditional volatility over time. So, the summary statistics of these tables are advantageous to control the risk or mitigate COVID cases' effect.

Conclusion

This paper discusses the daily rates of COVID-19 cases from four different countries, namely the United States, Spain, Italy, and China. The data shows a stochastic nature over time, so we estimated the stochastic volatility of daily rates. The daily rates of COVID-19 cases show the persistence behavior, meaning the movements of time series is correlated with their past observations and reflect stationarity at some past time lags. The persistence behavior was analyzed with the ARFIMA model parameter using Whittle likelihood (see Long memory test subsection). It is an effective method for reducing the number of MLE operations using the stationary property of COVID-19 data, and provides a good spectral approximation to the log-likelihood. In addition to this, the stochastic feature of stationary data helps to model the high fluctuations or high rate of COVID cases with much certainty.

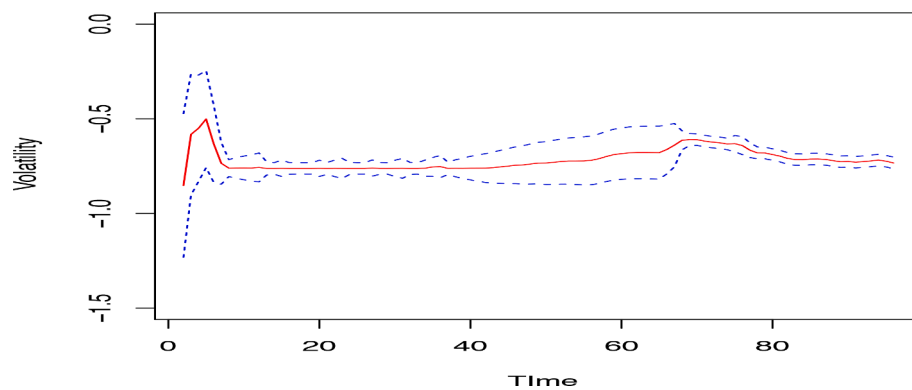In this study, we used the Kalman filtering technique in conjunction

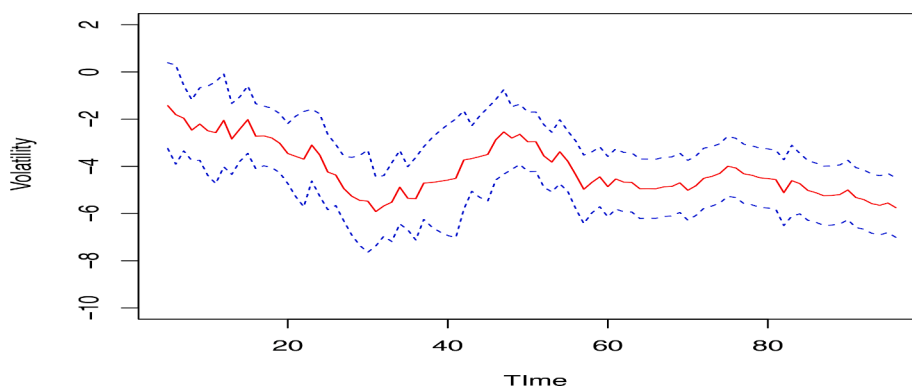**Fig. 6.** Sample path of one-step-ahead log-volatility, with ±3 standard prediction errors for USA COVID cases.

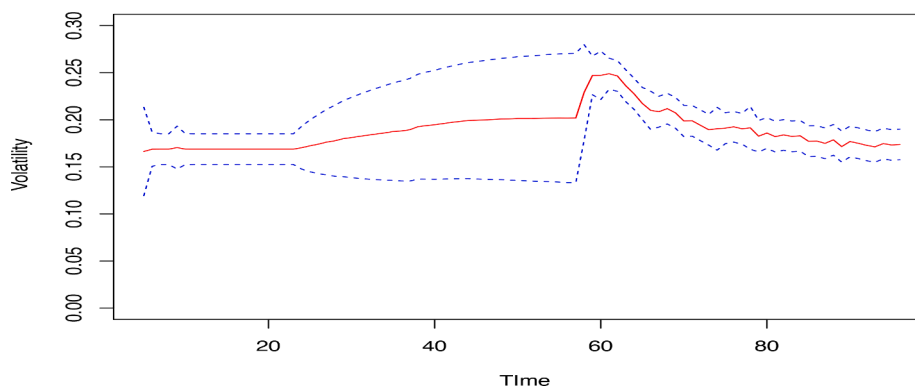**Fig. 7.** Sample path of one-step-ahead log-volatility, with ±3 standard prediction errors for China COVID cases.

Fig. 8. Sample path of one-step-ahead log-volatility, with ±3 standard prediction errors for Spain COVID cases.
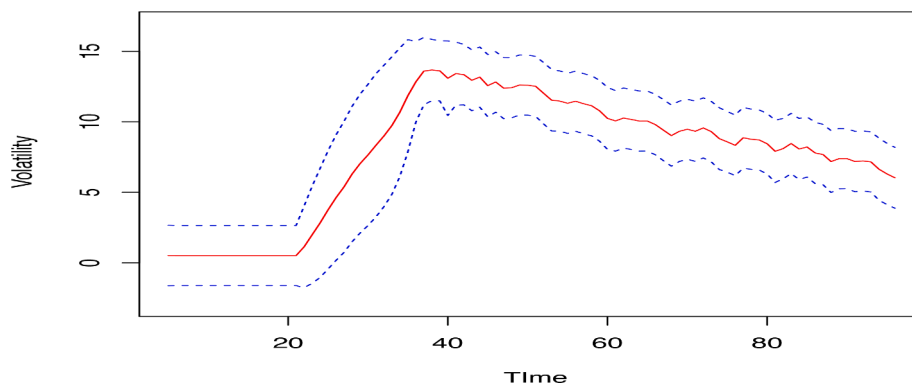
Fig. 9. Sample path of one-step-ahead log-volatility, with ±3 standard prediction errors for Italy COVID cases.

with the SV model for forecasting the data volatility. The process filters out the unnecessary information from data and provides the estimation of time-varying parameters that support non-constant conditional volatility. The results suggest that this stationary process forecast the volatility effectively with Kalman filtering. The one-step-ahead log-volatility with ± 3 standard prediction errors were shown over time (see Figs. 6–9), and the low errors of parameter estimation (see Tables 4–7) imply that the estimation is around the actual value. So the analysis is useful to detect the high rate of COVID-19 cases of a particular time. As we applied the test case for four leading countries of COVID cases, it can be applied to any country with a high new cases rate and new deaths rate. Although the vaccine for COVID-19 is now available, the number of cases is increasing. Therefore, detecting the high rate would allow us to raise awareness of self-protection and to take all the possible protective steps such as practicing social distancing, improving personal hygiene, covering the face with a mask, and other prescribed methods by experts.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Velavan TP, Meyer CG. The COVID-19 epidemic. Trop Med Int Health 2020;25(3): 278.

[2] Fauci AS, Lane HC, Redfield RR. Covid-19-navigating the uncharted 2020.

[3] Dhand R, Li J. Coughs and sneezes: their role in transmission of respiratory viral infections, including SARS-CoV-2. Am J Respir Crit Care Med 2020;202(5):651–9.

[4] Chowell G, Mizumoto K. The COVID-19 pandemic in the USA: what might we expect? The Lancet 2020;395(10230):1093–4.

[5] Khazanchi R, Beiter ER, Gondi S, Beckman AL, Bilinski A, Ganguli I. County-level association of social vulnerability with COVID-19 cases and deaths in the USA. J Gen Intern Med 2020;35(9):2784–7.

[6] https://datascience.nih.gov/covid-19-open-access-resources.

[7] Xu B, Gutierrez B, Mekaru S, Sewalk K, Goodwin L, Loskill A, Zarebski AE. Epidemiological data from the COVID-19 outbreak, real-time case information. Sci Data 2020;7(1):1–6.

[8] Hamzah FB, Lau C, Nazri H, Ligot DV, Lee G, Tan CL, Shaib MKBM. CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction. Bull World Health Organ 2020;1(32).

[9] Eliason SR. Maximum likelihood estimation-logic and practice, Quant Appl Soc Sci 193;96:1–10.

[10] Engle RF. Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation. Econometrica 1982;50(4):987–1007.

[11] Bollerslev T. Generalized autoregressive conditional heteroskedasticity. J Econometrics 1986;31:307–27.

[12] Bhuiyan MA Masum. Predicting Stochastic volatility for extreme fluctuations in high frequency time series, Open Access Theses & Dissertations. 2020. p. 2934.

[13] Janssen A, Drees H. A stochastic volatility model with flexible extremal dependence structure. Bernoulli 2016;22(3):1448–1490.

[14] Commandeur JF, Koopman SJ. An introduction to state space time series analysis. Oxford University Press 2007:107–21.

[15] https://www.sciencedirect.com/topics/engineering/gaussian-white-noise.

[16] Cipra T, Romera T. Robust Kalman filter and its application in time series analysis. Kybernetika 1991;27(6):481–94.

[17] Said SE, Dickey D. Testing for unit roots in autoregressive moving-average models with unknown order. Biometrika 1984;71:599–607.

[18] https://en.wikipedia.org/wiki/Autoregressivefractionallyintegratedmovingaverage.

[19] Kent JT, Mohammadzadeh M. Spectral approximation to the likelihood for an intrinsic gaussian random field. J Multivariat Anal 1997;70:136–55.

[20] https://www.sciencedirect.com/topics/mathematics/bernoulli-distribution.

[21] Efstathios P, Dimitris NP. The asymptotic size and power of the augmented Dickey-Fuller test for a unit root. Economet Rev 2018.

[22] Chatterjee K, Chatterjee K, Kumar A, Shankar S. Healthcare impact of COVID-19 epidemic in India: A stochastic mathematical model. Med J Armed Forces India 2020;76(2):147-155.