Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Time series modelling to forecast the confirmed and recovered cases of COVID-19

Mohsen Maleki\textsuperscript{a}, Mohammad Reza Mahmoudi\textsuperscript{b,c,*}, Darren Wraith\textsuperscript{d}, Kim-Hung Pho\textsuperscript{e}

\textsuperscript{a} Department of Statistics, University of Isfahan, Isfahan, Iran
\textsuperscript{b} Institute of Research and Development, Duy Tan University, Da Nang, 550000, Vietnam
\textsuperscript{c} Department of Statistics, Faculty of Science, Fasa University, Fasa, Fars, Iran
\textsuperscript{d} Institute of Health and Biomedical Innovation (IHBI), Queensland University of Technology (QUT), Queensland, Australia
\textsuperscript{e} Fractional Calculus, Optimization and Algebra Research Group, Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

\textbf{ARTICLE INFO}

\textbf{Keywords:}
Coronaviruses
COVID-19
Prediction
Autoregressive model
Two pieces distributions based on the scale mixtures normal distribution

\textbf{ABSTRACT}

Coronaviruses are enveloped RNA viruses from the Coronaviridae family affecting neurological, gastrointestinal, hepatic and respiratory systems. In late 2019 a new member of this family belonging to the Betacoronavirus genera (referred to as COVID-19) originated and spread quickly across the world calling for strict containment plans and policies. In most countries in the world, the outbreak of the disease has been serious and the number of confirmed COVID-19 cases has increased daily, while, fortunately the recovered COVID-19 cases have also increased. Clearly, forecasting the “confirmed” and “recovered” COVID-19 cases helps planning to control the disease and plan for utilization of health care resources. Time series models based on statistical methodology are useful to model time-indexed data and for forecasting. Autoregressive time series models based on two-piece scale mixture normal distributions, called TP-SMN–AR models, is a flexible family of models involving many classical symmetric/asymmetric and light/heavy tailed autoregressive models. In this paper, we use this family of models to analyze the real world time series data of confirmed and recovered COVID-19 cases.

1. Introduction

Coronaviridae family includes two main subfamilies Coronavirinae and Torovirinae. The member genera include Alphacoronavirus, Betacoronavirus, Gammacoronavirus, Torovirus, and Bafnavirus. They are a huge family of viruses that affect neurological, gastrointestinal, hepatic and respiratory systems and can be grown among humans, bats, mice, livestock, birds, and others \cite{1–3}. In the Coronaviridae family, a well-known type of virus called SARS coronavirus (SARS-CoV) distributed from animal to animal and humans \cite{4}. Another type of coronavirus, called MERS coronavirus (MERS -CoV), significantly distributed from human to human in 2012 \cite{4}. In 2019 many cases in China with respiratory diseases were reported by the World Health Organization (WHO), with evidence that these cases originated from a seafood market in Wuhan \cite{5}. In 2019 a new type of virus called COVID-19 (novel coronavirus, 2019-nCoV), belonging to the Betacoronavirus genera of Coronaviridae family, spread from Wuhan in China \cite{6}.

Evidence that COVID-19 is distributed from human to human has been verified by the Centers for Disease Control and Prevention (CDC), and also reported that COVID-19 is spreading by touching surfaces, close contact, air, or objects that contain viral particles. The incubation period of COVID-19 is at least 14 days \cite{7}, and it can spread to others in the incubation period. Finally note that the incubation period and median age of confirmed cases are respectively 3 days and 47.0 years \cite{8}.

Preparation and controlling the outbreak of COVID-19 diseases requires thorough planning and policies. Some researchers have used statistical and mathematical modelling. In China, the number of unreported COVID-19 cases has been mathematically estimated in Refs. 9. Also based on the information of some Japanese passengers in Wuhan \cite{10}, estimated the rate of the infection for COVID-19 in Wuhan. The results indicated a rate of 9.5% for infection and a rate from 0.3% to 0.6%, for death. Based on mathematical modelling in Ref. [11], the transmission risk of COVID-19 is on average about 6.47 per person and predicted the time that the peak of COVID-19 will be reached. Estimation of a sustained human-to-human transmission equal to 0.4 for COVID-19 using the information of 47 patients has been done in Ref.

\* Corresponding author. Institute of Research and Development, Duy Tan University, Da Nang, 550000, Vietnam.

E-mail addresses: m.maleki.stat@gmail.com (M. Maleki), mohammadreamahmoudi@duytan.edu.vn, mahmoudi.m.r@fasau.ac.ir (M.R. Mahmoudi), d.wraith@qut.edu.au (D. Wraith), phokimhung@tdtu.edu.vn (K.-H. Pho).

https://doi.org/10.1016/j.tmaid.2020.101742

Received 9 March 2020; Received in revised form 7 May 2020; Accepted 9 May 2020
Available online 13 May 2020

1477-8939/ © 2020 Elsevier Ltd. All rights reserved.
The modelling, estimation and prediction of the prevalence of viruses and the epidemiological characteristics are important issues in providing the equipment needed to cope with their consequences. Forecasting of the cases and transmission risk of West Nile virus (WNV) has been provided by Ref. [14]. For further modelling and forecasting of the spread of several viruses such as the hepatitis A virus, Ebola, SARS, influenza A and MERS, refer to Refs. [15–21]. To have a suitable plan for COVID-19, forecasting the future confirmed cases are critical. An optimization method, named FPASSA-ANFIS, has been proposed by Ref. [22] to model the number of confirmed cases of COVID-19 and to predict its future values using collected data in China. Forecasting various data about COVID-19, by mathematical and statistical models, is very important to a program of cutting the transmission chain of diseases; see, e.g., Ref. [23–26].

According to credible daily reports from the World Health Organization and other world-renowned institutions in the field of public health, the total number of COVID-19 confirmed cases has increased in different countries, especially in U.S.A, Italy, Spain and Iran. Although the spread of COVID-19 has many dangers, fortunately reports show that the total number of COVID-19 recovered cases has also increased. Increasing the number of recovered cases, along with reducing or stabilizing the number of confirmed cases is important to control the spread of the COVID-19 and leads to stability of the rate of infections in the world. So modelling and forecasting the numbers of confirmed and recovered COVID-19 cases has an important role to plan the control of the spread of the COVID-19 in the world. Cumulative numbers of the confirmed and recovered COVID-19 cases, which are reported daily by the proposed organizations, on each day depend on their values on the past days. So using autoregressive time series model can be a useful tool to model, analyze and forecast the confirmed and recovered cases of COVID-19. The SIR epidemic modelling can be done at local (country) level but the autoregressive model can be good to look at overall patterns. The autoregressive time series model is a flexible tool to model dependent data and has been used to estimate and forecast many real practical problems, see Refs. [27–34]. In fact, the autoregressive model, determines the probabilistic behavior of the current values $X_t$ based on a linear combination of past values $\{X_{t-1}, X_{t-2}, X_{t-3}, \ldots\}$, in the form of:

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + Z_t; \quad t = 0, \pm 1, \pm 2, \ldots, \quad (1)$$

where the error terms $\{Z_t\}$ are generally assumed to be uncorrelated and identically probabilistically distributed random variables from a distribution, and denoted by $\{X_t\}$-AR($p$). See e.g. Ref. [39, 40].

Because in many real world time series data, classical modelling based on the symmetrical/light-tailed distributions are not satisfactory, in our methodology we have used autoregressive time series model (1) based on asymmetric/heavy-tailed $TP$-SMN distributions. Therefore we assume the error terms $Z_t$ in (1) are distributed as $TP$-SMN distributions, denoted by $Z_t$-$TP$-SMN($\mu$, $\sigma$, $\nu$, $\gamma$) and $\{X_t\}$-$TP$-SMN-$AR(p)$. See details of the proposed distributions and model in Ref. [34–38].

In this paper we modeled the total number of confirmed and recovered COVID-19 cases in the world by the proposed autoregressive time series model so-called $TP$-SMN-$AR$ models which includes the symmetric Gaussian and asymmetric heavy-tailed non-Gaussian autoregressive time series models. The various members of the proposed autoregressive models were fitted initially to the historical numbers of confirmed and recovered COVID-19 cases in the world. Then, the autoregressive time series that has the best fit to each of the dataset is selected. Finally, the selected models are used to predict the number of confirmed and recovered COVID-19 cases in the world from 21-Apr-2020 up to 30-Apr-2020, and we measure the differences between the real and predicted values to show the performance of the models. Therefore, the main contribution points of the current study are as follows: an improved autoregressive time series model based on the $TP$-SMN distributions, and a new efficient predictive model applied to predict and estimate the confirmed and recovered COVID-19 cases in the world using past and current data. Note that a sample copy of the code is available from the authors upon request.

2. Modelling the confirmed and recovered cases of the COVID-19 in the world

The coronavirus (COVID-19) is affecting about 212 countries and territories around the world and two international conveyances. The daily data for COVID-19 in the world are reported by the China National Health Commission (NHC) and World Health Organization (WHO). In this part we fit the $TP$-SMN-$AR$ time series model to the total confirmed COVID-19 cases from 22-Jan-2020 to 30-Apr-2020 and also to the total recovered COVID-19 cases from 02-Feb-2020 to 30-Apr-2020, in the world.

Time series plots of the total confirmed and recovered cases are plotted in Fig. 1 and Fig. 2 respectively. The proposed time series plots are not stationary because they are increasing and show signs of a trend. After some suitable transformations described in Ref. [40], we
obtain stationary data. Also using model selection criteria [34,39,40] the best TP–SMN–AR models (the autoregressive models based on the two-piece t distributions) were fitted to the stationary series of the confirmed and recovered cases and are given by.

- The confirmed COVID-19 cases; TP–SMN–AR(7) model:
  \[ X_t = -0.9399X_{t-1} - 1.0438X_{t-2} - 1.1067X_{t-3} - 0.9825X_{t-4} - 0.9364X_{t-5} - 0.8105X_{t-6} - 0.3623X_{t-7} + Z_t, \]
  where \( Z_t \sim TP-T(\mu = -103.3424, \sigma = 4084.3481, \gamma = 0.4559, \nu = 2.1000). \)

- The recovered COVID-19 cases; TP–SMN–AR(2) model:
  \[ X_t = -0.5865X_{t-1} - 0.2020X_{t-2} + Z_t, \]
  where \( Z_t \sim TP-T(\mu = -0.0048, \sigma = 0.0250, \gamma = 0.4756, \nu = 3.1146). \)

The histograms of the estimated errors (residuals) based on the estimated heavy-tailed TP–SMN densities are superimposed in Fig. 3 and show the suitable performance of the estimated models to the stationary series of total confirmed and recovered COVID-19 cases datasets. Also the auto-correlation function (ACF) plots of the residuals presented in Fig. 4 show the suitability of the fitted models.

To further demonstrate the goodness of fit of the model, we eliminated the last 10 days of the confirmed and recovered cases (2020-Apr-21 to 2020-Apr-30), and then fitted the TP–SMN–AR models and provided forecasts. Table 1 contains the predictions and 98% confidence intervals for this analysis. Also Fig. 5, Fig. 6 and Fig. 7, show the forecasted values which are superimposed on the plots of the real values of the confirmed and recovered COVID-19 cases in the world.

To evaluate the accuracy of the predictions, we use the mean relative percentage error (MAPE), which for the confirmed COVID-19 cases is 0.22% and for the recovered COVID-19 cases is 1.6% which are reasonably low values demonstrating the suitability of the proposed models for prediction.

Finally note that the proposed TP–SMN–AR models include as special or limiting cases the more standard autoregressive time series models used in the literature. In particular, some model selection criteria such as Akaike information criteria (AIC), Bayesian information criteria (BIC), and Box–Pierce and Ljung–Box tests on the residuals, demonstrate that the proposed fitted TP–SMN–AR models are more reasonable than other well-known counterparts.
3. Conclusion

Coronaviruses are a huge family of viruses that affect neurological, gastrointestinal, hepatic, and respiratory systems. The number of confirmed cases has increased daily in different countries, especially in the U.S.A., Italy, Spain, Iran, China and others. The spread of COVID-19 has many dangers and needs strict special plans and policies. Therefore, to consider plans and policies, predicting and forecasting the future confirmed and recoveries cases are critical. The autoregressive time series models are a useful tool to model data over time. However, some of the standard time series models are based on the assumption that the error term or residuals are symmetric (Gaussian). There exist many situations in the real world that the assumption of symmetric distribution of the error terms is not satisfactory. In our methodology, we considered autoregressive time series models based on the two-piece scale mixture normal (TP-SMN) distributions. The results indicated that the proposed method performed well in forecasting confirmed and recovered COVID-19 cases in the world. Using model selection criteria, the proposed models were also more reasonable than the standard Gaussian autoregressive time series model which is the simplest member of our proposed models. For future works, we suggest that the researchers apply cyclostationary, almost cyclostationary and simple processes [41–47] based on the TP-SMN distributions, instead of stationary...
Funding
No funding

CRediT authorship contribution statement

Mohsen Maleki: Data curation, Validation, Writing - original draft.
Mohammad Reza Mahmoudi: Conceptualization, Methodology, Software, Supervision.
Darren Wraith: Visualization, Investigation, Writing - review & editing.
Kim-Hung Pho: Visualization, Investigation, Writing - review & editing.

Declaration of competing interest

The authors declare no conflict of interest.

References

[1] Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. J Med Virol 2020. https://doi.org/10.1002/jmv.25681.
[2] Ge XY, Li JI, Yang XL, Chmura AA, Zhu G, Epstein JH, Mazet JK, Hu B, Zhang W, Peng C, et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature 2013;503:535–8.
[3] Wang LF, Shi Z, Zhang S, Field H, Dazak P, Eaton BT. Review of bats and SARS. Emerg Infect Dis 2006;12:1834.
[4] Cauchemez S, Van Kerkhove M, Riley S, Donnelly C, Fraser C, Ferguson N. Transmission scenarios for Middle East Respiratory Syndrome Coronavirus (MERS-CoV) and how to tell them apart. EuroSurveill. Bull. Eur. Sur Les Mal. Transm. Eur. Commun. Dis. Bull. 2013;18:20502.
[5] Organization WH. Novel coronavirus (2019-nCoV) 2020. 2020 Available online:https://www.who.int/, (accessed on 27 January 2020).
[6] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020;395:565–74.
[7] Cheng ZJ, Shan J. Novel coronavirus: where we are and what we know. Infection 2019;2020. https://doi.org/10.1007/s15010-020-01401-y.
[8] Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui DS, et al. Clinical characteristics of 2019 novel coronavirus infection in China. medRxiv 2020. doi:10.1101/2020.02.06.20020974 Available online:https://www.medrxiv.org/content/early/2020/02/09/2020.02.06.20020974.full.pdf (accessed on 9 February 2020).
[9] Zhao S, Musa SS, Lin Q, Ran J, Yang G, Wang W, Lou Y, Yang L, Gao D, He D, et al. Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of January 2020: a data-driven modelling analysis of the early outbreak. J Clin Med 2020;9:388.
[10] Nishiura H, Kobayashi T, Yang Y, Hayashi K, Miyama T, Kinoshita R, Linton NM, Jung Sm, Yuan B, Suzuki A, et al. The rate of underascertainment of novel coronavirus (2019-nCoV) infection: estimation using Japanese passengers data on evacuation flights. J Clin Med 2020;9:419.
transmission risk of the 2019-nCoV and its implication for public health interventions. J Clin Med 2020;9:462.

[12] Thompson RN. Novel coronavirus outbreak in wuhan, China, 2020: intense surveillance is vital for preventing sustained transmission in new locations. J Clin Med 2020;9:498.

[13] Jung SM, Akhmetzhanov AR, Hayashi K, Linton NM, Yang Y, Yuan B, Kobayashi T, Kinoshita R, Nichiura H. Real-time estimation of the risk of death from novel coronavirus (2019-nCoV) infection: inference using exported cases. J Clin Med 2020;9:523.

[14] DeFelice NB, Little E, Campbell SR, Shaman J. Ensemble forecast of human West Nile virus cases and mosquito infection rates. Nat Commun 2017;8:1–6.

[15] Ture M, Kurt I. Comparison of four different time series methods to forecast hepatitis A virus infection. Expert Syst Appl 2006;31:41–6.

[16] Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. Proc Natl Acad Sci USA 2012;109:20425–30. J. Clin. Med. 2020, 9, 674 14 of 15.

[17] Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012–2013 season. Nat Commun 2013;4:1–10.

[18] Shaman J, Yang W, Kandula S. Inference and forecast of the current west african Ebola outbreak in Guinea, Sierra Leone and Liberia. PLoS Curr 2014.6. https://doi.org/10.1371/currents.outbreaks.3408774290b1af952dd7caed877c6b8ff.

[19] Box George, Jenkins Gwilym M, Reinsel Gregory C. Time series analysis: forecasting and control. third ed. Prentice-Hall0130607746; 1994.

[20] Moravveji M, Khodadadi Z, Maleki M. A bayesian analysis of two-piece distributions based on the scale mixtures of normal family. Iran J Sci Technol A (Sciences) 2019;43(3):991–1001.

[21] Maleki M, Mahmoudi MR, Contreras-Reyes JE. The Skew-Reflected-Gompertz distribution for analyzing symmetric and asymmetric data. J Comput Appl Math 2019;349:132–41.

[22] Box George, Jenkins Gwilym M, Reinsel Gregory C. Time series analysis: forecasting and control. third ed. Prentice-Hall0130607746; 1994.

[23] Brockwell PJ, Davis RA. Time series: theory and methods. second ed. New York: Springer; 2009.

[24] Maleki M, Mahmoudi MR, Pak A. Testing the difference between two independent time series models. Iran J Stat Technol A (Sciences) 2017;41:665–9.

[25] Maleki M, Heydari MH, Avazzadeh Z. On the asymptotic distribution for the periodograms of almost periodically correlated (cyclostationary) processes. Digit Signal Process 2020;96:102597.

[26] Maleki M, Barkhordar Z, Khodadadi Z, Wraith D. A robust class of homoscedastic processes based on the unrestricted skew normal process. J Stat Comput Simulat 2019;89(1):38–51.