Pak. J. Statist.
2025 Vol. 41(3), 187-206

STATISTICAL MODELING AND PREDICTION OF PANDEMIC USING CLASSICAL AND MACHINE LEARNING APPROACH

Ahtasham Gul[1], Faridoon Khan[2], Muhammad Farooq[3], Amjad Javaid[1§], Zawar Hussain[4] and **Bassam ElZaghmouri[5]**

[1] Pakistan Bureau of Statistics, Islamabad, Pakistan.
Email: ahtashamgul@gmail.com; amjadsandhu@hotmail.com

[2] Department of Creative Technologies, Faculty of Computing and Artificial Intelligence, Air University, Islamabad, Pakistan.
Email: faridoon.khan@au.edu.pk

[3] Department of Statistics, COMSATS University Islamabad Lahore Campus, Pakistan. Email: muhammadfarooq@cuilahore.gov.pk

[4] Department of Statistics, Faculty of Computing, The Islamia University of Bahawalpur, Pakistan. Email: zhlangah@yahoo.com

[5] Department of Computer Science, Al-Ahliyya Amman University, Amman, Jordan. Email: b.elzaghmouri@ammanu.edu.jo

[§] Corresponding author

ABSTRACT

A variant of truncated distribution known as Gumbel-Truncated exponential distribution (G-TEXPD) is applied to estimate the fatality charge of sufferers from COVID-19 in Pakistan using classical and machine learning algorithms. In the classical approach, maximum likelihood estimation (MLE) technique is implemented to estimate the parameters of G-TEXPD. For predictive modeling, authors have taken into consideration the machine learning (ML) algorithms and some statistical models. To evaluate the prognostication functioning, commonly known accuracy measures such as mean absolute error (MAE), root mean square error (RMSE) and mean absolute percent error (MAPE) are applied. The results indicate that machine learning algorithms are highly effective at predicting the mortality rate. These outputs are compared with baseline statistical models.

KEYWORDS

$T-X_{tr}$ family of distributions, truncation, Gumbel-Truncated exponential distribution, maximum likelihood estimation, machine learning, extrapolation.

Mathematics Subject Classification (2010):
60E05, 60E10, 62E15, 62E17, 62P30.

1. INTRODUCTION

With the surge in patients' number, many countries are confronting a shortage of medical supplies, because of their high demand. However, knowing the spread level of a disease and forecasting its trend accurately is one of the most important steps, which can assist in

© 2025 Pakistan Journal of Statistics

curbing the spread of a pandemic, particularly, in the most populous countries like China, Pakistan, India, etc. A reliable forecast of the pandemic like COVID-19 is very useful for the government to allocate resources more effectively.

It is observed that finite data sets are estimated and predicted by using statistical models having infinite domain in real life. For instance, Ristić and Balakrishnan (2012) studied Gamma-exponentiated exponential (GEE) distribution to model the life of Guinea pigs receiving tubercle bacilli intake. Hossam et al. (2022) introduced new version of Gumbel distribution called Novel alpha power Gumbel distribution to make inferences of two different COVID-19 data sets. Fayomi et al. (2022) proposed exponentiated Gumbel-G (EGuG) family of distributions to study the meteorological and medical data sets. Cordeiro et al. (2012) proposed Kumaraswamy-Gumbel (KumGu) to analyze the hydrological data set. Similarly, Sindhu et al. (2021) proposed Exponentiated transformation of Gumbel Type-II (ETGT-II) for modeling the mortality cases due to COVID-19. The aforementioned probability functions have a domain from $(0, \infty)$ contrary to the data set which is defined on a bounded range.

The concept of continuous distributions for univariate is originated by Pearson (1895) using differential equations. Subsequently, Johnson (1949) employed the method of transformation to develop continuous distributions. Tukey (1960) introduced the quantile function based method. Lee et al. (2013) reviewed the literature and pointed out that since 1980, bulk of the generated families of distributions are developed by using the compounding of distributions or induction of additional parameter(s) into a conventional distribution. Likewise, Log-Gamma-generated families and Exponentiated Half-Logistic family of distributions are introduced by Amini et al. (2014) and Cordeiro et al. (2014) respectively.

Eugene et al. (2002) suggested the idea of integral transformation to establish a family of Beta generated distributions. The CDF and PDF of random variable Z defined by Eugene is

$$F(z) = \int_0^{G(z)} s\,(w)dw, \tag{1}$$

where w is Beta random variable and $g(z)$ follows any probability distribution and PDF is

$$f(z) = \frac{1}{B(\theta, \epsilon)} g(z)G^{\theta 1}(z)\big(1 - G(z)\big)^{\epsilon 1}, 0 \le z \le 1. \tag{2}$$

Nadarajah and Kotz (2004, 2006) proposed Beta-Gumbel and Beta-Exponential distributions and Akinsete Akinsete et al. (2008) established Beta-Pareto distribution by using Eugene's methodology. Cordeiro & de Castro (2011) and Jones (2009) extended the idea of Eugene by substituting the Kumaraswamy distribution (1980) in place of Beta distribution. The PDF of Generalized Kumaraswamy distribution is given by

$$g(z) = \theta\beta f(z)F^{\theta-1}(z)\left(1 - F^{\theta}(z)\right)^{\beta-1}, \qquad z \in (0,1). \tag{3}$$

The aforementioned methods suggested by Eugene et al. (2002), Jones (2009) and Cordeiro & de Castro (2011) constrained the proposed distributions defined over the $(0,1)$ and baseline distributions remain fixed.

Alzaatreh et al. (2013) improved the idea of Eugene et al. (2002) by substituting the PDF of any continuous variable X in the place of PDF of Beta random variable. The new proposed family of distribution having CDF as:

$$F(z) = \int_0^{W(G(z))} s\,(x)dx, \qquad l < z < u \tag{4}$$

having PDF

$$f(z) = \left\{\frac{d}{\mathrm{dz}}\mathrm{w}\big(\mathrm{G(z)}\big)\right\}\,\mathrm{s(w(G(z)))}, \qquad \mathrm{l < z < u.} \tag{5}$$

Tahir et al. (2016) utilized the technique of Alzaatreh et al. (2013) to develop the Logistic-X family of probability densities. Similarly Alzaghal et al. (2013) introduced a new generator to generate Exponentiated $T - X$ family of probability distributions, i.e.

$$W\big(G(z)\big) = -\log(1 - G^c(z)). \tag{6}$$

In real life, we deal with data having finite domain rather than infinite domain such as medicine, epidemiology, biometry, engineering, hydrology, environmental sciences and reliability analysis. For instance, to study survival problems, we desire to study survival time after taking cardiovascular attack handling. The cardiovascular attack indication is the beginning time of our study. Similarly, in engineering, the lifespan of a car tyre after use of 50,000 km. Thus, probability models having infinite domain reduced to finite domain are called the truncated distributions. Chapman (1956) is the pioneer to estimate the parameters of a truncated Gamma distribution. Broeder & Gerard (1955) and Hegde & Dahiya (1989) suggested different methods to estimate the parameters of truncated distributions. Jawitz (2004) highlighted the importance of truncated distributions in hydrology by comparing truncated moment expressions (TMEs) with complete moments using six standard distributions. The author demonstrated that complete moments are special case of general TMEs (for more, see Jawitz (2004)).

Many researchers in the literature have forecasted the COVID-19 spread using various statistical tools namely, Ceylan (2020), Yousaf et al. (2020), Chintalapudi et al. (2020), Satrio et al. (2021) and Benvenuto et al. (2020) have adopted the pure statistical models. On the other hand, Chakraborty and Ghosh (2020) mixed the wavelet-based forecasting (WBF) and Autoregressive integrated moving average (ARIMA) techniques to propose a novel approach. Similarly, several ML algorithms and hybrid procedures are implemented by Chimmula & Zhang (2020), Wu et al. (2020), Tomar and Gupta (2020), Roosa et al. (2020), Santosh (2020), Pourhomayoun & Shakibi (2021), Sah et al. (2022), Ahmad et al. (2022) for the prediction of COVID-19 spread.

This study evaluates the forecasting performance of penalization techniques, namely ALasso and Enet, in comparison with the traditional time series model known as SETAR model. To the best of our knowledge, no existing study in the literature has compared updated penalization techniques with the SETAR model on COVID-19 data. Therefore, this study presents the first comparative analysis of penalization methods and the SETAR model using COVID-19 data from Pakistan. The afore-mentioned techniques are delineated in the following subsections.

Generally, statistical learning has two basic objectives: exploring the relevant predictors and ensuring high precision in forecasting. Feature selection plays a decisive role when the basic true model has a sparse depiction. The forecasting performance of the estimated model can be improved by identifying the crucial predictors. Using the following objective function, we can minimize the squared residuals using

$$\frac{1}{2} \parallel \Re_t - \Re_{t-i} m \parallel^2 + \sum_{p=1}^{k} v_p \, (\mid P \mid), \tag{7}$$

where penalty function is represented by $m = (m_1, m_2, \ldots . m_k)$ and $V_p(|.|)$, which takes different forms for various procedures. Equation (7) comprises of two parts: the foremost part indicates the squared sum of residual and gauge the model goodness. The second part represents the penalty term, which exhibits the complexity of model. The regularization methods are explained as:

1.1 Least Absolute Shrinkage and Selection Operator

If the penalty function in Equation (7) is replaced by L_1 norm, then the function tends to Lasso Tibshirani (2011). In general, the Lasso approach is capable to force the inactive covariates to zero and merely keep the active covariates. If the set of covariates are highly associated with each other, then Lasso picks a single covariate of them. Besides, the Lasso is biased in the selection of covariates (Khan et al. 2022).

1.2 Adaptive Lasso

The Lasso was updated by adding the weighting scheme i.e., $c_p = \mid m_p \mid^{(-\tau)}$ in the Lasso penalty function, named as Adaptive Lasso (ALasso). As a result, the ALasso is unbiased in the covariates selection (Zou, 2006).

1.3 Elastic Net

If the penalty in Equation (7) is replaced by ridge penalty (L_2) i.e., $V_{p(|\bullet|)} = z|m|^2$, the resultant estimator leads to ridge estimator. The ridge estimator just enforces the coefficients near zero and thereby it just allays the problem of multicollinearity and fails to perform covariates selection. If both penalties, namely L_1 and L_2 are added in the equation then the estimator tends to elastic net (Enet). The Enet selects the entire group of covariates provided if it is associated with each other (Zou and Hastie (2005), Khan and Albalawi (2024)). Hyperparameter tuning for Lasso and Elastic Net is done with k-fold cross-validation on the training set. In particular, we employed 10-fold cross-validation to determine the best values of the hyperparameters as followed by Khan et al. (2024).

1.4 SETAR Model

Unlike the ML algorithms described in the earlier subsections, the SETAR model is a part of conventional statistical methods of time series. Tong and Lim (2009) are the first who proposed it as the extension of the autoregressive model. The SETAR model is basically a nonlinear by nature, parametric and a particular case of Markov switching models. It assumes that the variable nature changes, once it enters into a different regime. Equation (8) contains some parameters, including m (number of regimes), ß (threshold indicator) and a (lagged realizations). Moreover, the low and high regimes are indicated by mL and mH respectively. Moreover, the low and high regimes are indicated by mL and mH respectively. The model SETAR $(m = 3, a)$ can be illustrated as:

$$\Re_t = 1_{(\beta_{(t-1\leq th)})}(l_0^{mL} + l_1^{mL}(\Re_{t-1}) + \cdots + l_a^{mL}(\Re_{t-a})) + 1((\beta_{(t-1>th)}))(l_0^{mH} + l_1^{mH}(\Re_{t-1}) + \cdots + l_a^{mH}(\Re_{t-a}). \quad (8)$$

In case of fixed β, the model tends to be linear and is estimated through conditional least squares.

In the recent years, estimation of model parameters is shifted from analytical approaches such as maximum likelihood (ML) and least square estimation (LSE) to data driven approaches. One of most popular technique for parameter estimation which is free of likelihood function is approximate Bayesian computation (ABC) established by Beaumont et al. (2002). Sisson et al. (2018) have discussed ABC in detail especially for inference in several fields like, genetics, climate, ecology and nuclear imaging.

Now a day, machine learning is widely applied in predictive analytics, image recognition, finance, medical science etc. Among many machine learning techniques, neural network has a lot of attention by researcher in almost every field particularly for forecasting and prediction and data driven parameter estimation. The traces of neural point estimation begins from Chon et al. (1997) who used artificial neural network to estimate parameters of linear and non-linear ARIMA model. However, the use of neural networks has become notable in recent years as demonstrated by the works of Creel (2017), Flagel et al. (2019), Rudi et al. (2022), Banesh et al. (2021), Lenzi et al. (2023), Gerber and Nychka (2021), Zammit-Mangion et al. (2020) and Gaskin et al. (2023) in the fields of stock returns, population genetics, time series, spatial fields, spatio-temporal fields and agent-based models, respectively. Abbasi et al. (2008) employed neural network to estimate the parameters of Weibull distribution.

The main motivation of writing this manuscript is to model COVID-19 data having finite domain using a truncated probability distribution to avoid the extrapolating issues. Traditionally, statistical models with infinite domain have been employed to analyze finite datasets. However, this often leads to extrapolation beyond the observed data range, which can increase uncertainty in estimation, prediction and interpretation. To overcome this limitation, the current study employs the Type-2 Gumbel-Truncated Exponential Distribution (G-TEXPD) introduced by Gul et al. (2023) to model COVID-19 mortality rates (per million population) in Pakistan. The G-TEXPD is particularly suitable for this context due to its strong capability to model right-skewed, heavy-tailed data within a finite domain commonly observed in pandemic-related datasets such as COVID-19 mortality rates. Moreover, improved forecasting can help assess the effectiveness of implemented policies in reducing the burden on the healthcare system. It also enables the government to evaluate mitigation plans and synchronize policies based on region-specific forecasts.

The remaining sections of the manuscript are arranged as follows: In Section 2, G-TEXPD introduced by Gul et al. (2023) is defined. In Section 3, model parameters of Type-2 Gumbel-Truncated Exponential distribution (G-TEXPD) are estimated by using MLE method. In Section 4, the feasibility of proposed model is studied by fitting it to COVID-19 data set in Pakistan and comparing it with some other baseline models. The post sample forecasting of COVID-19 data is accomplished in Section 5. Finally, some concluding remarks and way forward are summarized in Section 6.

2. TYPE-2 GUMBEL-TRUNCATED EXPONENTIAL DISTRIBUTION (G-TEXPD)

Gul et al. (2023) defined a new technique to establish a family of truncated distributions called $T - X_{tr}$ family of distributions by using log inverse truncated survival function, i.e.

$$W(G(X_{tr})) = -\log\{1 - G(x_{tr}|x_{tr} > \tau)\}. \tag{9}$$

Let T be a random variate with probability density function $s(t)$ and cumulative distribution function $S(t)$ on interval $[-\infty \leq a \leq t \leq b \leq \infty)$.

Also X_{tr} be a left-truncated non-negative random variable with PDF $g(X_{tr})$ and CDF $G(X_{tr})$ on domain $[\tau, \infty)$.

So the CDF of transformer-transformed truncated family $(T - X_{tr})$ of distributions is

$$F(X_{tr}) = \int_{\tau}^{-\log\{1-G(x_{tr}|x_{tr} > \tau)\}} s(t)dt, \tag{10}$$

$$F(X_{tr}) = S[-\log\{1 - F(X_{tr})\}] \tag{11}$$

The respective PDF of transformer-transformed truncated family $T - X_{tr}$ of distributions is

$$f(X_{tr}) = s[-\log\{1 - G(x_{tr}|x_{tr} > \tau)\}]\frac{g(x_{tr}|x_{tr} > \tau)}{1 - G(x_{tr}|x_{tr} > \tau)}, \quad x_{tr} > \tau. \tag{12}$$

$$f(X_{tr}) = s[-\log\{1 - G(X_{tr})\}]\left\{\frac{g(X_{tr})}{1 - G(X_{tr})}\right\}, \tag{13}$$

$$f(X_{tr}) = s[H(X_{tr})]h[X_{tr}]. \tag{14}$$

Suppose X_{tr} be a random variate following truncated Exponential distribution having PDF

$$f(X_{tr}) = \theta\, s\{\theta(x_{tr} - \tau)\}, I(\xi, \infty), \tau > 0 \tag{15}$$

and CDF

$$F(X_{tr}) = S\big(\theta(x_{tr} - \tau)\big). \tag{16}$$

where τ is the truncation point.

Let T be the type-2 Gumbel random variable having two parameters i.e. α and β, then the PDF is

$$g(t) = \alpha\beta t^{-\alpha-1}\, e^{-\beta t^{-\alpha}}, \qquad t > 0, \tag{17}$$

and CDF

$$G(t) = e^{-\beta t^{-\alpha}}. \tag{18}$$

Using Equation (15), the resultant distribution Gumble type 2 -truncated exponential distribution (G-TEXPD) is

$$G(X_{tr}) = \alpha\beta\theta^{-\alpha}e^{-\beta\theta^{-\alpha}(x-a)^{-\alpha}}(x-a)^{-\alpha-1}, \qquad x > a. \tag{19}$$

The CDF of resultant distribution is

$$G(X_{tr}) = e^{-\beta\theta^{-\alpha}(x-a)^{-\alpha}} \tag{20}$$

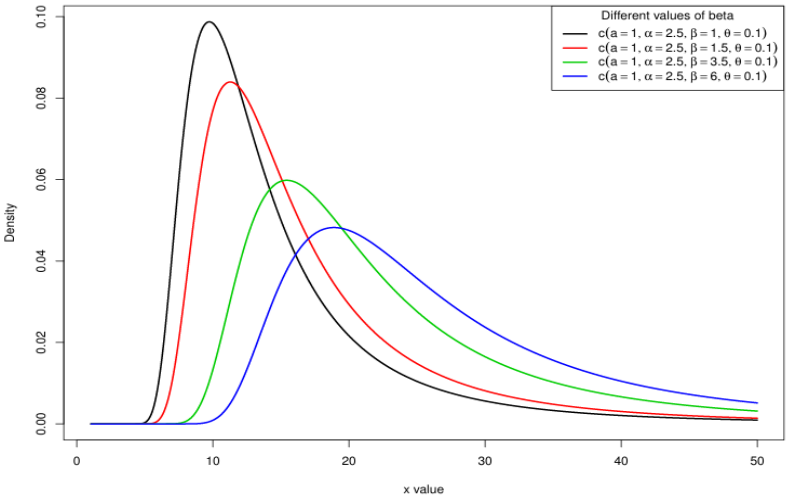

Figure 1(a): Different Values of β

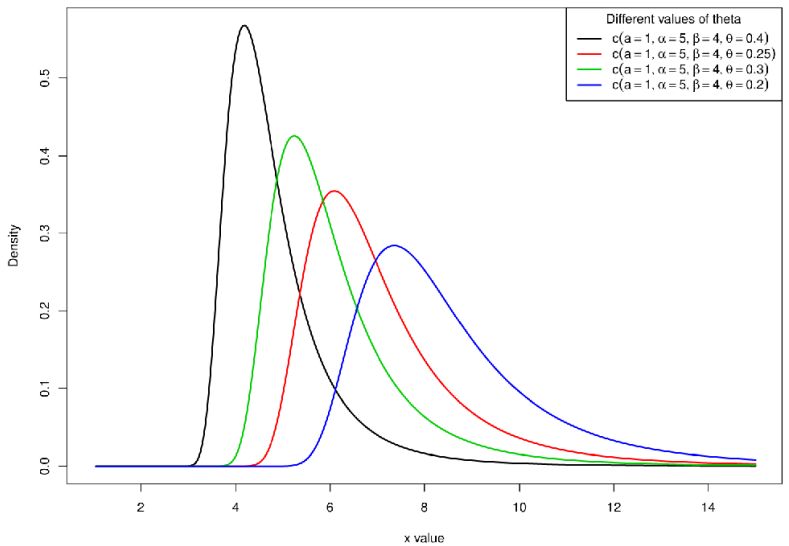

Figure 1(b): Different Values of θ

Figure 1: The pdf of the G-TEXPD for Different Values of α, β and θ

Figure 1 demonstrates distinct aspects of G − TEXPD for various combination of values of parameters. The above Figure 1(a) is described at various values of β reflecting that increase in β, cause the shape to tend to heavy tail (leptokurtic) i.e. it will exclude or mitigate the extreme values. The additional parameter θ in Figure 1(b) portray the mesokurtic shape of G − TEXPD with an increase in θ.

The rth raw moment of G-TEXPD is defined as:

$$E(X_{tr}^{r}) = \frac{\beta^{r/\alpha}}{\theta^{r}} \sum_{k=0}^{r} \binom{r}{k} \left(\frac{\theta a}{\beta^{1/\alpha}}\right)^{r-k} \left|\overline{1-\frac{k}{\alpha}}\right. \tag{21}$$

3. PARAMETERS ESTIMATION OF PROPOSED MODEL USING CLASSICAL APPROACH

Let X be distributed as $G - TEXPD$ having PDF (Equation 19), then its log-likelihood function is given by:

$$L(\underline{\theta}; X_{tr}) = n\log\alpha + n\ln\beta - n\alpha\log\theta - \alpha\sum_{j=1}^{n}\log(x_j - a) - \sum_{j=1}^{n}\log(x_j - a) - \frac{\beta}{\theta^{\alpha}}\sum_{j=1}^{n}\log(x_j - a)^{-\alpha}, \tag{22}$$

where

$$a = Min[x_j]. \tag{23}$$

$$\frac{L(a,\alpha,\beta,\theta;X_{tr})}{\partial\alpha} = \begin{bmatrix} \frac{n}{\alpha} - \mathrm{n}\log(\theta) - \sum_{j=1}^{n}\log\left(x_j - a\right) \\ -\frac{\beta}{\theta^{\alpha}}\log(\theta)\sum_{j=1}^{n}\ln\left(x_j - a\right)^{-\alpha} \\ -\frac{\beta}{\theta^{\alpha}}\sum_{j=1}^{n}(x_j - a)^{-\alpha}\log\left(x_j - a\right)) \end{bmatrix} = 0. \tag{24}$$

$$\frac{L(a,\alpha,\beta,\theta;x_T)}{\partial\beta} = \frac{n}{\beta} - \frac{\sum_{j=1}^{n}(x_j - a)^{-\alpha}}{\theta^{\alpha}} = 0. \tag{25}$$

$$\frac{L(a,\alpha,\beta,\theta;x_T)}{\partial\theta} = \frac{\alpha\beta}{\theta^{\alpha+1}}\sum_{j=1}^{n}(x_j - a)^{-\alpha} - \frac{n\alpha}{\theta} = 0. \tag{26}$$

4. APPLICATION OF PROPOSED MODEL ON REAL LIFE DATA: COVID-19 DATA

The competence of the Gumbel-Truncated Exponential distribution is established by modeling the mortality rate (per million population) from Corona virus disease 2019 (COVID-19) in Pakistan downloaded from Corona virus (Covid-19) Pandemic Statistics and Research (https://github.com/owid/covid19data). This pandemic immobilized the everyday life of well-beings particularly, it has created public health catastrophe at global level. The data set comprising of 135 observations is recorded as:

16.852 16.874 16.967 17.043 17.060 17.211 17.247 17.282 17.327 17.358 17.380
17.451 17.487 17.522 17.522 17.553 17.647 17.695 17.775 17.780 17.793 17.882
17.940 17.957 18.037 18.051 18.068 18.082 18.095 18.135 18.148 18.246 18.264
18.290 18.353 18.646 18.681 18.708 18.788 18.877 19.032 19.174 19.196 19.267
19.272 19.396 19.418 19.600 19.729 19.782 19.840 19.853 19.925 19.969 19.996
20.093 20.147 20.218 20.355 20.511 20.631 20.693 20.853 20.959 20.968 20.972
20.995 21.021 21.070 21.123 21.168 21.252 21.319 21.425 21.563 21.572 21.909
21.954 21.980 21.985 22.025 22.083 22.087 22.096 22.291 22.318 22.353 22.425
22.700 22.820 22.877 23.073 23.242 23.499 23.588 23.655 23.663 23.792 23.819
23.912 23.956 24.179 24.298 24.334 24.418 24.916 25.138 25.906 25.928 25.977
26.008 26.066 26.234 26.852 27.207 27.251 28.206 28.406 28.419 28.739 29.019
29.041 29.325 30.231 31.297 31.949 33.477 33.686 34.094 34.312 35.337 35.360
36.337 37.722 53.610.

Table 1
Descriptive Statistics of Mortality Rate (per Million Population) from COVID-19 in Pakistan

Min.	Max.	Q1	Q2	Q3	Mean	Var.	Skewness	Kurtosis
16.85	53.61	18.32	21.02	24.07	22.43	29.86	2.15	10.25

The descriptive statistics of mortality rate from COVID-19 in Pakistan placed in Table 1 shows that the data is positive skewed and heavy tailed to right side (platy kurtic).

Table 2
Negative Log-Likelihood Function, MLEs of Model Parameters, the Corresponding SEs (in parentheses) and the Statistics AIC and BIC for Mortality Rate (per Million Population) from COVID-19 in Pakistan

Model	Estimates					Statistic	
	$\hat{l}$	$\hat{a}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\theta}$	AIC	BIC
G-TEXPD	**-393.044**	**16.85**	**0.868**	**0.218**	**0.083**	**792.089**	**800.805**
			(0.052)	(0.737)	(0.321)		
Weibull	-432.251	—	24.559	3.706	—	868.502	874.312
			(0.203)	(0.607)			
Gamma	-403.766	—	0.936	21.002	—	811.533	817.313
			(0.114)	(2.536)			
Exponential	-554.893	—	0.045	—	—	1111787	1114.69
			(0.004)				

Table 3
Goodness of Fit Statistic of Mortality Rate (per million) for COVID-19 Patients in Pakistan

Model	G-TEXPD	Weibull	Gamma	Exponential
K-Smirnov	**0.13**	0.219	0.123	0.528
C-Von	**0.684**	1.376	0.636	8.556
A-Darling	**0.146**	8.184	3.994	40.003

Table 2 provides the estimates of model parameters by using ML method while Table 3 gives different statistics to check the goodness of fit for mortality rate (per million population) from COVID-19 in Pakistan along with the comparison of G-TEXPD with Weibull, Gamma and exponential distributions. The results in Table 2 and 3 connote that G-TEXPD better estimates the mortality rate of COVID-19 in Pakistan. Thus, G-TEXPD can be recommended as the best model based on AIC, BIC and goodness of fit test contrary to other models and we can utilize it as an alternative probability algorithm in prediction, particularly with reference to Pandemic data.

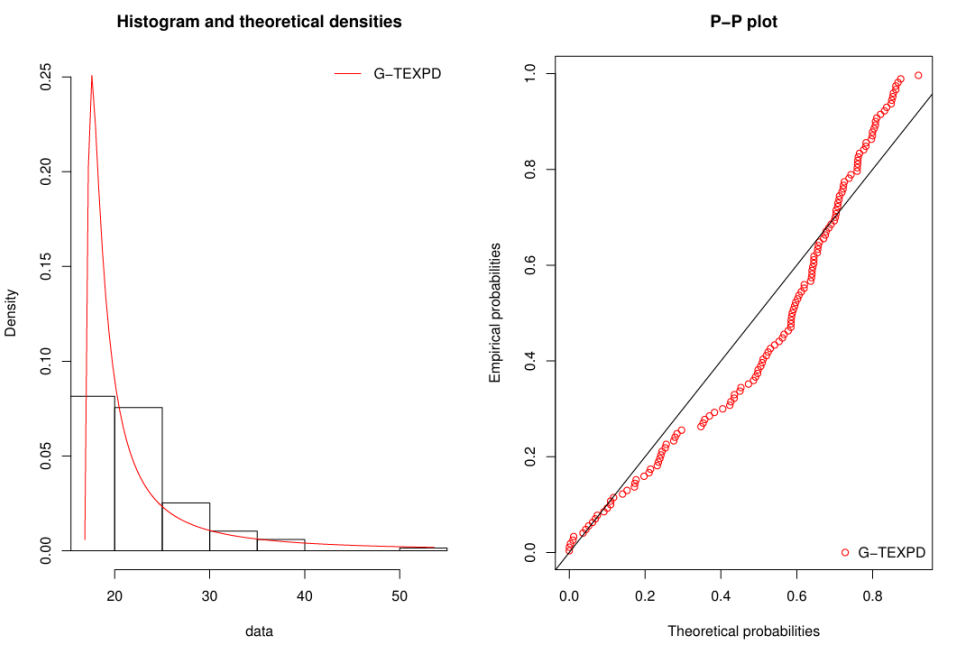

Figure 2: The Fitted PDF of Gumbel-TEXPD on Mortality Rate (per million Population) from COVID-19 in Pakistan along with its Probability Plot

In Figure 2, the observed probabilities corresponding to predicted probabilities are plotted which depict better fitting of the $G-TEXPD$ to the mortality rate (per million population) from COVID-19 in Pakistan.

5. POST SAMPLE COVID-19 FORECASTING

The objective of this section is to determine which model better fits the COVID-19 data as discussed in the earlier section. In this context, the penalization procedures (ALasso and Enet) are compared with the SETAR model via the widely used two statistical measures i.e., root mean square error (RMSE) and mean absolute error (MAE). The mathematical expressions can be illustrated as:

$$MAE = mean(|\ \Re_t - \hat{\Re}_t\ |) \tag{27}$$

$$RMSE = \sqrt{mean(\Re_t - \hat{\Re}_t)^2} \tag{28}$$

The $\Re_t$ and $\hat{\Re}_t$ are representing the observed and predicted values of COVID-19 data in Equations (27 and 28) respectively.

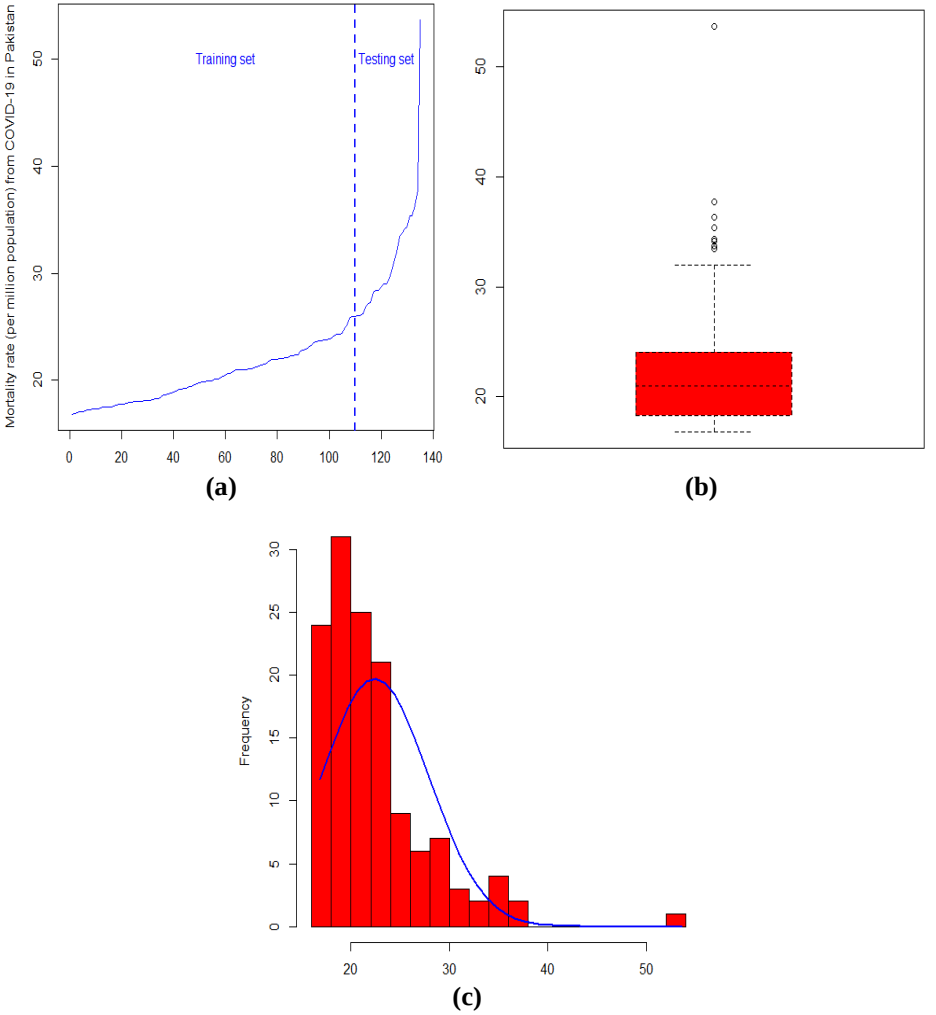

Figure 3: Visual Representation of the Mortality Rate Data of COVID-19 Patients

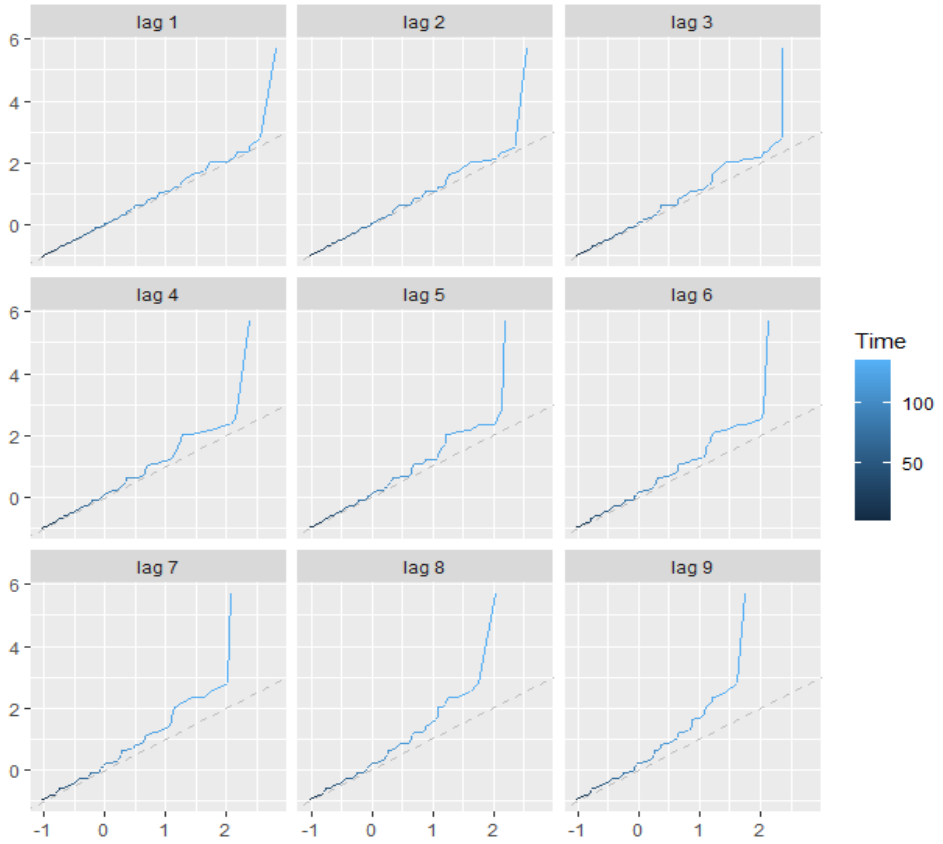

Figure 4: Correlation between Outcome Variable and Lagged Variable

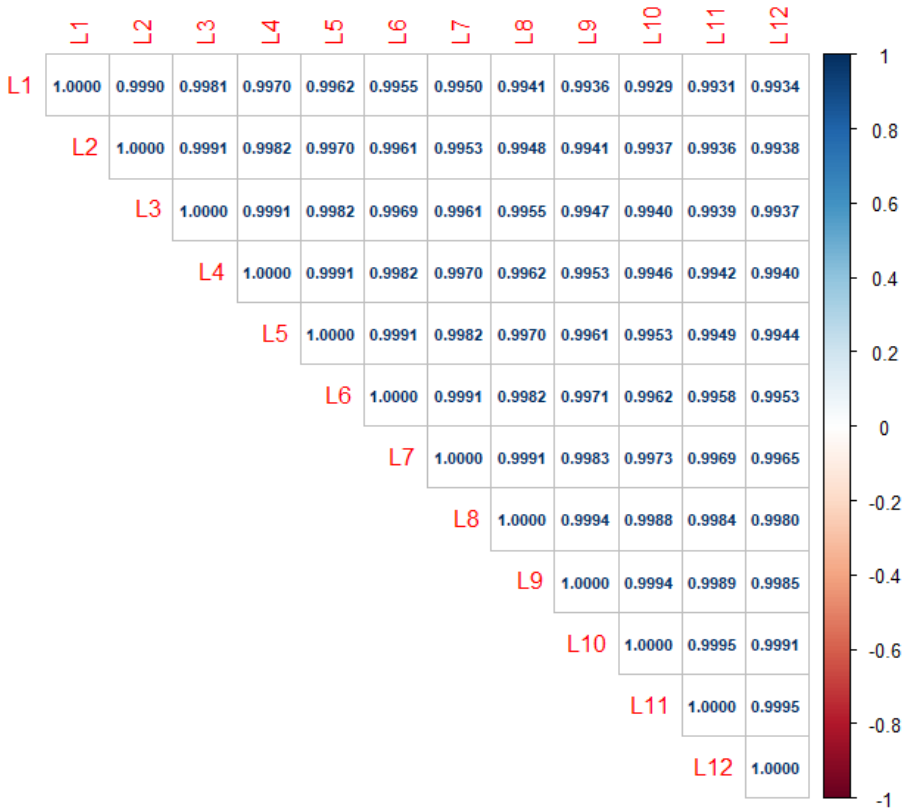

Figure 5: Correlation between Lagged Variables

In Figure 4, it can be observed that how the outcome variable is associated with their own lags. To be more specific, the effect decreases as we go into the recent past $(\Re_{t-1}, R_{t-2}, \dots)$ from the R_t. Therefore, in time series analysis the lagged realizations plays a key role. Similarly, the pairwise correlation among the lagged variables is plotted in Figure 5. We can notice that there exists an extreme correlation among the set of predictors (lagged variables). In econometrics and statistics, such a phenomenon is known as "the severe case of multicollinearity" which is very harmful to a model. Under such circumstances, the implementation of ordinary least square method is not useful as the standard errors of the regression coefficients are invalid and thereby the output of statistical tests is fallacious.

Hence, multicollinearity is an issue of serious concern in linear regression analysis. To resolve this problem, there are numerous tools available in the literature, including Principal Components Regression (PCR), Partial Least Square (PLS), Ridge Regression (RR), Least absolute shrinkage and selection operator (Lasso), smoothly clipped absolute deviation, Adaptive Lasso (ALasso), elastic net (Enet) etc. Among the early mentioned list, ALasso and Enet are the most well-known and updated versions to tackle the multicollinearity problem.

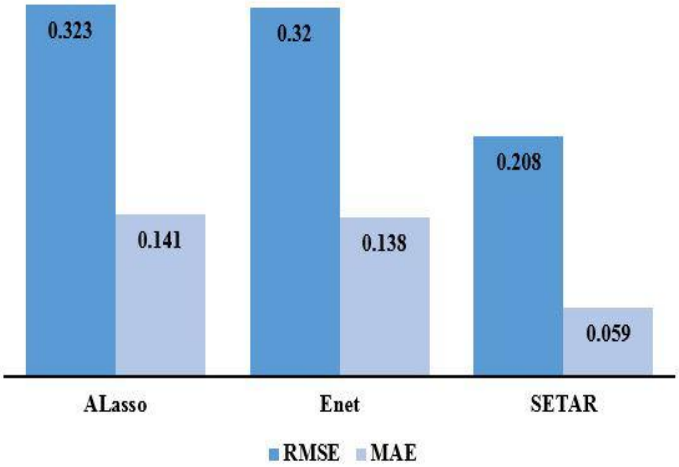

Figure 6: Comparison Flowchart

The accuracy measures such as RMSE and MAE are computed using the COVID-19 pandemic data (mortality rate) taken from Pakistan sketched in Figure 6. The RMSE and MAE computed for SETAR model are 0.059 and 0.208, whereas for ALasso and Enet, the RMSE and MAE are (0.323 and 0.141) and (0.32 and 0.138) respectively. It can be inferred that the loss functions computed for SETAR model are substantially smaller than the competing approaches. Furthermore, the Enet produces improved forecasts as compare to ALasso.

6. CONCLUDING REMARKS

This study addresses a key limitation in traditional statistical modeling that is the use of distributions with infinite domain to analyze real world datasets that inherently lie within the finite domains. This leads to extrapolation of data that significantly increases uncertainty in estimation, prediction, and interpretation. To overcome this limitation, the Type-2 Gumbel-Truncated Exponential Distribution ($G-TEXPD$) is applied to model COVID-19 mortality rate (per million population) in Pakistan. The results indicate that G-TEXPD provides a better fit to pandemic-related data compared to conventional models such as the Weibull, Gamma, and exponential distributions. Statistical results demonstrate that truncated models for finite datasets significantly enhance the precision of estimation, prediction and interpretation, particularly for skewed and heavy tail data. These findings confirm its effectiveness as a flexible and reliable framework for modeling real world, finite, and skewed pandemic-related data.

Beyond distributional modeling, this study also investigates forecasting performance by comparing penalization methods (ALasso and Enet) with SETAR model using the same COVID-19 mortality rate data. The results indicate that SETAR model outperforms

contrary to penalization methods in terms of forecasting. Moreover, Elastic Net demonstrated superior forecasting performance compared to Adaptive Lasso among the penalization methods. As we observed that the distribution of COVID-19 data is right skewed due to outlying data points in the dataset. Thus, because of nonlinearity, the SETAR model shows the superior performance in forecasting contrary to penalizations methods, which are linear. This study employed a relatively small dataset, the size of which might lower the generalization of results. Moreover, analysis was limited to univariate techniques only. These factors may restrict the capability to examine sophisticated relations among variables. Furthermore, the dataset was cleaned to preserve real world characteristics and imputation methods are not applied. The overlay plots are sketched to compare observed and predicted trends to illustrate model performance more intuitively.

In future research, we are motivated to develop bivariate or multivariate version of G-TEXPD to study the Pandemic problems in different dimensions. Moreover, employing Bayesian estimation techniques and validate the model on datasets from different regions.

ACKNOWLEDGEMENT

The authors would like to thank anonymous reviewers for their constructive comments and suggestions that helped to improve the manuscript significantly.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interests regarding the publication of this manuscript.

REFERENCES

1. Abbasi, B., Rabelo, L. and Hosseinkouchack, M. (2008). Estimating parameters of the three-parameter Weibull distribution using a neural network. *European Journal of Industrial Engineering*, 2(4), 428-445.
2. Ahmad, Z., Almaspoor, Z., Khan, F. and El-Morshedy, M. (2022). On predictive modeling using a new flexible weibull distribution and machine learning approach: Analyzing the covid-19 data. *Mathematics*, 10(11), 1792.
3. Akinsete, A., Famoye, F. and Lee, C. (2008). The beta-pareto distribution. *Statistics*, 42(6), 547-563.
4. Alzaatreh, A., Lee, C. and Famoye, F. (2013). A new method for generating families of continuous distributions. *Metron*, 71(1), 63-79.
5. Alzaghal, A., Famoye, F. and Lee, C. (2013). Exponentiated t-x family of distributions with some applications. *International Journal of Statistics and Probability*, 2(3), 31-49.
6. Amini, M., MirMostafaee, S. and Ahmadi, J. (2014). Log-gamma-generated families of distributions. *Statistics*, 48(4), 913-932.
7. Banesh, D., Panda, N., Biswas, A., Van Roekel, L., Oyen, D., Urban, N., Grosskopf, M., Wolfe, J. and Lawrence, E. (2021). Fast gaussian process estimation for large-scale

in situ inference using convolutional neural networks. In 2021 *IEEE International Conference on Big Data (Big Data)*, pages 3731-3739. IEEE.
8. Beaumont, M.A., Zhang, W. and Balding, D.J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4), 2025-2035.
9. Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S. and Ciccozzi, M. (2020). Application of the arima model on the covid-2019 epidemic dataset. *Data in brief*, 29, 105340.
10. Ceylan, Z. (2020). Estimation of covid-19 prevalence in Italy, Spain, and France. *Science of the Total Environment*, 729, 138817.
11. Chakraborty, T. and Ghosh, I. (2020). Real-time forecasts and risk assessment of novel coronavirus (covid-19) cases: A data-driven analysis. *Chaos, Solitons & Fractals*, 135, 109850.
12. Chapman, D.G. (1956). Estimating the parameters of a truncated gamma distribution. *The Annals of Mathematical Statistics*, 27(2), 498-506.
13. Chimmula, V.K.R. and Zhang, L. (2020). Time series forecasting of covid-19 transmission in canada using lstm networks. *Chaos, Solitons & Fractals*, 135, 109864.
14. Chintalapudi, N., Battineni, G. and Amenta, F. (2020). Covid-19 virus outbreak forecasting of registered and recovered cases after sixty-day lockdown in Italy: A data driven model approach. *Journal of Microbiology, Immunology and Infection*, 53(3), 396-403.
15. Chon, K.H. and Cohen, R.J. (1997). Linear and nonlinear arma model parameter estimation using an artificial neural network. *IEEE Transactions on Biomedical Engineering*, 44(3), 168-174.
16. Cordeiro, G.M., Alizadeh, M. and Ortega, E.M. (2014). The exponentiated half-logistic family of distributions: Properties and applications. *Journal of Probability and Statistics*, 2014(1), 864396.
17. Cordeiro, G.M. and de Castro, M. (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, 81(7), 883-898.
18. Cordeiro, G.M., Pescim, R.R. and Ortega, E.M. (2012). The kumaraswamy generalized half-normal distribution for skewed positive data. *Journal of Data Science*, 10(2), 195-224.
19. Creel, M. (2017). Neural nets for indirect inference. *Econometrics and Statistics*, 2, 36-49.
20. Den Broeder Jr, G.G. (1955). On parameter estimation for truncated Pearson, type iii distributions. *The Annals of Mathematical Statistics*, 26(4), 659-663.
21. Eugene, N., Lee, C. and Famoye, F. (2002). Beta-normal distribution and its applications. *Communications in Statistics-Theory and Methods*, 31(4), 497-512.
22. Fayomi, A., Khan, S., Tahir, M.H., Algarni, A., Jamal, F. and Abu-Shanab, R. (2022). A new extended Gumble distribution: Properties and application. *Plos one*, 17(5), e0267142.
23. Flagel, L., Brandvain, Y. and Schrider, D.R. (2019). The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Molecular Biology and Evolution*, 36(2), 220-238.

24. Gaskin, T., Pavliotis, G.A. and Girolami, M. (2023). Neural parameter calibration for large-scale multiagent models. *Proceedings of the National Academy of Sciences*, 120(7), e2216415120.
25. Gerber, F. and Nychka, D. (2021). Fast covariance parameter estimation of spatial gaussian process models using neural networks. *Stat*, 10(1), e382.
26. Gul, A., Mohsin, M., and Almaspoor, Z. (2023). A new family of distributions to encounter the extrapolating issues. *Mathematical Problems in Engineering*, 2023(1), 9397398.
27. Hegde, L.M. and Dahiya, R.C. (1989). Estimation of the parameters of a truncated gamma distribution. *Communications in Statistics-Theory and Methods*, 18(2), 561-577.
28. Hossam, E., Abdulrahman, A.T., Gemeay, A.M., Alshammari, N., Alshawarbeh, E. and Mashaqbah, N.K. (2022). A novel extension of Gumble distribution: Statistical inference with covid-19 application. *Alexandria Engineering Journal*, 61(11), 8823-8842.
29. Jawitz, J.W. (2004). Moments of truncated continuous univariate distributions. *Advances in Water Resources*, 27(3), 269-281.
30. Johnson, N.L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1-2), 149-176.
31. Jones, M. (2009). Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, 6(1), 70-81.
32. Khan, F. and Albalawi, O. (2024). Analysis of fat big data using factor models and penalization techniques: A monte carlo simulation and application. *Axioms*, 13(7), 418.
33. Khan, F., Muhammadullah, S., Sharif, A. and Lee, C.-C. (2024). The role of green energy stock market in forecasting china's crude oil market: An application of its approach and sparse regression models. *Energy Economics*, 130, 107269.
34. Khan, F., Urooj, A., Khan, S.A., Khosa, S.K., Muhammadullah, S. and Almaspoor, Z. (2022). Evaluating the performance of feature selection methods using huge big data: A Monte Carlo simulation approach. *Mathematical Problems in Engineering*, 2022(1), 6607330.
35. Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46(1), 79-88.
36. Lee, C., Famoye, F. and Alzaatreh, A.Y. (2013). Methods for generating families of univariate continuous distributions in the recent decades. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(3), 219-238.
37. Lenzi, A., Bessac, J., Rudi, J. and Stein, M.L. (2023). Neural networks for parameter estimation in intractable models. *Computational Statistics & Data Analysis*, 185, 107762.
38. Nadarajah, S. and Kotz, S. (2004). The beta Gumble distribution. *Mathematical Problems in Engineering*, 2004(4), 323-332.
39. Nadarajah, S. and Kotz, S. (2006). The beta exponential distribution. *Reliability Engineering & System Safety*, 91(6), 689-697.

40. Pearson, K. (1895). X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London* (A), 186, 343-414.
41. Pourhomayoun, M. and Shakibi, M. (2021). Predicting mortality risk in patients with covid-19 using machine learning to help medical decision-making. *Smart Health*, 20, 100178.
42. Ristić, M.M. and Balakrishnan, N. (2012). The gamma-exponentiated exponential distribution. *Journal of Statistical Computation and Simulation*, 82(8), 1191-1206.
43. Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J.M., Yan, P. and Chowell, G. (2020). Real-time forecasts of the covid-19 epidemic in china from February 5 to February 24 2020. *Infectious Disease Modelling*, 5, 256-263.
44. Rudi, J., Bessac, J. and Lenzi, A. (2022). Parameter estimation with dense and convolutional neural networks applied to the Fitzhugh-nagumo ode. In *Mathematical and Scientific Machine Learning*, pages 781-808. PMLR.
45. Sah, S., Surendiran, B., Dhanalakshmi, R., Mohanty, S.N., Alenezi, F. and Polat, K. (2022). Forecasting covid-19 pandemic using prophet, ARIMA, and hybrid stacked lstmgru models in India. *Computational and Mathematical Methods in Medicine*, 2022(1), 1556025.
46. Santosh, K. (2020). Covid-19 prediction models and unexploited data. *Journal of Medical Systems*, 44(9), 1-4.
47. Satrio, C.B.A., Darmawan, W., Nadia, B.U. and Hanafiah, N. (2021). Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and prophet. *Procedia Computer Science*, 179, 524-532.
48. Sindhu, T.N., Shafiq, A. and Al-Mdallal, Q.M. (2021). Exponentiated transformation of Gumble type-ii distribution for modeling covid-19 data. *Alexandria Engineering Journal*, 60(1), 671-689.
49. Sisson, S.A., Fan, Y. and Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. CRC Press.
50. Tahir, M., Cordeiro, G.M., Alzaatreh, A., Mansoor, M. and Zubair, M. (2016). The logistic-X family of distributions and its applications. *Communications in Statistics, Theory and Methods*, 45(24), 7326-7349.
51. Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), 273-282.
52. Tomar, A. and Gupta, N. (2020). Prediction for the spread of covid-19 in India and effectiveness of preventive measures. *Science of the Total Environment*, 728, 138762.
53. Tong, H. and Lim, K.S. (2009). Threshold autoregression, limit cycles and cyclical data. In *Exploration of A Nonlinear World: An Appreciation of Howell Tong's Contributions to Statistics*, pages 9-56. World Scientific.
54. Tukey, J.W. (1960). The practical relationship between the common transformations of percentages of counts and of amounts. *Statistical Techniques Research Group Technical Report*, 36.

55. Wu, K., Darcet, D., Wang, Q. and Sornette, D. (2020). Generalized logistic growth modeling of the covid-19 outbreak: comparing the dynamics in the 29 provinces in china and in the rest of the world. *Nonlinear Dynamics*, 101(3), 1561-1581.
56. Yousaf, M., Zahir, S., Riaz, M., Hussain, S.M. and Shah, K. (2020). Statistical analysis of forecasting covid-19 for upcoming month in Pakistan. *Chaos, Solitons & Fractals*, 138, 109926.
57. Zammit-Mangion, A. and Wikle, C.K. (2020). Deep integro-difference equation models for spatio-temporal forecasting. *Spatial Statistics*, 37, 100408.
58. Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.
59. Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.