Journal of Data Science 18 (3), 483–495
July 2020
DOI: 10.6339/JDS.202007_18(3).0015

COVID-19 Fatality: A Cross-Sectional Study using Adaptive Lasso Penalized Sliced Inverse Regression

Kaida Cai[2], Wenqing He[1], and Grace Y. Yi[*1,2]

[1]*Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada*

[2]*Department of Computer Science, University of Western Ontario, London, Ontario, Canada*

Abstract

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronvirus, which was declared as a global pandemic by the World Health Organization on March 11, 2020. In this work, we conduct a cross-sectional study to investigate how the infection fatality rate (IFR) of COVID-19 may be associated with possible geographical or demographical features of the infected population. We employ a multiple index model in combination with sliced inverse regression to facilitate the relationship between the IFR and possible risk factors. To select associated features for the infection fatality rate, we utilize an adaptive Lasso penalized sliced inverse regression method, which achieves variable selection and sufficient dimension reduction simultaneously with unimportant features removed automatically. We apply the proposed method to conduct a cross-sectional study for the COVID-19 data obtained from two time points of the outbreak.

Keywords *coronavirus disease 2019; infection fatality rate; multiple index model; risk factors*

1 Introduction

Since January 2020, many regions in China have experienced an outbreak of the coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus (SARS-Cov-2). Researchers from different fields have been studying epidemiological and clinical characteristics of the COVID-19. For example, Epidemiology Working Group for NCIP Epidemic Response (2020) studied epidemiological characteristics of the COVID-19 in China. Xu et al. (2020) investigated pathological characteristics of a patient who died from severe infection with SARS-CoV-2. Zhang et al. (2020) accessed how livers would be affected using the available case studies. A number of researchers studied clinical features of individual patients of COVID-19 (Guan et al., 2020; Huang et al., 2020; Chen et al., 2020; Wang et al., 2020d). Investigations on the characteristics of COVID-19 by using statistical and machine learning methods were also available. For instance, Hu et al. (2020) employed a modified stacked auto-encoder for modeling transmission dynamics of the epidemics to forecast confirmed cases of COVID-19 across China. Wang et al. (2020b) developed an epidemiological forecast model with an R package to assess the intervention effect on the COVID-19 epidemic within and outside Hubei in China.

As COVID-19 becomes a global pandemic, one of the greatest concerns is to identify the fatality risk for infected populations which usually differ in multiple features. It is important to understand what features of the populations are associated with the high fatality risk. At the individual level, some work was available to describe the relationship between patient level

*Corresponding author. Email: gyi5@uwo.edu.

Received April, 2020; Accepted July, 2020

features and the fatality risk factors. For instance, Ji et al. (2020) investigated the potential association between the COVID-19 fatality and health-care resources. Shi et al. (2020) explored the association between the cardiac injury and the fatality of the patients with the COVID-19. Clinical Features of 69 cases with COVID-19 were studied by Wang et al. (2020d). Different risk factors that may be associated with the infected populations with COVID-19 were considered by different researchers, including temperature (Ma et al., 2020; Wang et al., 2020c), humidity (Wang et al., 2020c), age structure (Onder et al., 2020), and the cardiovascular disease (Clerkin et al., 2020). While each of the possible risk factors has been separately studied to reveal the potential association with COVID-19, it is unclear how those factors may interactively affect the fatality of the COVID-19. It is critical to examine how the COVID-19 fatality may be associated with potential risk factors as a group instead of on an individual basis. To this end, in this paper we employ the multiple index model to facilitate the relationship between possible risk factors and the fatality rate of COVID-19. Such a model is advantageous in its flexibility of accommodating various kinds of association with an unspecified model function.

To reduce the variable dimension of the multiple index model, we employ the sufficient dimension reduction method, a powerful tool for reducing the variable dimension of the multiple index model which can be employed to identify important risk factors by obtaining the central subspace (Cook, 1998). To estimate the central subspace of the sufficient dimension reduction, many methods have been developed, including the most widely used sliced inverse regression (SIR) (Li, 1991), the sliced average variance estimation (SAVE) method (Cook and Weisberg, 1991), the principal Hessian directions (pHd) method (Li, 1992), and the iterative Hessian transformation method (Cook et al., 2002). The shrinkage SIR based on the Lasso penalty for sufficient dimension reduction was proposed by Ni et al. (2005). Li and Nachtsheim (2006) and Li (2007) produced sparse estimates of the basis for the central subspace by combining a regression-type formulation with the Lasso penalty. Other penalized sufficient dimension reduction methods include the constrained canonical correlation procedure (Zhou et al., 2008), the SCAD max penalized SIR method (Wu and Li, 2011), and the Lasso-SIR (Lin et al., 2019).

Motivated by Zou (2006) and Lin et al. (2019), we propose an adaptive Lasso penalized sliced inverse regression method for the multiple index model to identify the possible risk factors for the infection fatality rate of COVID-19. The proposed method develops a model-free variable selection procedure which does not require the specification of a parametric model for the underlying true process. It estimates the central subspace of the multiple index model and selects the important features simultaneously.

The remainder of this article is organized as follows. Section 2 describes the COVID-19 data sets to be analyzed. Section 3 presents the proposed method and the implementation algorithm. The analysis results are reported in Section 4. Section 5 concludes the article with a discussion.

2 Descriptions of the COVID-19 Data

To interpret the fatality risk of infected populations, we use the infection fatality rate (IFR) in percent of COVID-19, defined as

$$\text{Infection Fatality Rate (in Percent)} = \frac{\text{Total Number of Deaths}}{\text{Total Number of Confirmed Cases}} \times 100.$$

We consider the data from $n = 64$ developed or developing countries in Asia, Europe, Africa, America, and Oceania. The Asian countries include Armenia, Azerbaijan, Bahrain, Iran, Israel, Japan, Kazakhstan, Malaysia, Pakistan, Philippines, Qatar, Saudi Arabia, Singapore,

Table 1: Dictionary of COVID-19 data covariates.

Feature	Variable	Data Source	Web Link (https://)
Age structure	x_1	World Bank	data.worldbank.org
Smoking prevalence	x_2	World Bank	data.worldbank.org
PM2.5 air pollution	x_3	World Bank	data.worldbank.org
Cardiovascular disease	x_4	Global Health Data Exchange	ghdx.healthdata.org
Chronic respiratory diseases	x_5	Global Health Data Exchange	ghdx.healthdata.org
Physicians	x_6	Our World in Data	ourworldindata.org/coronavirus-data
Hospital beds	x_7	Our World in Data	ourworldindata.org/coronavirus-data
Blood pressure	x_8	Risk Factor Collaboration	ncdrisc.org/data-downloads.html
Average temperature	x_9	Weather Base	www.weatherbase.com
Average relative humidity	x_{10}	Weather Base	www.weatherbase.com
Number of serious cases	x_{11}	Worldometers	www.worldometers.info/coronavirus
Number of tests	x_{12}	Worldometers	www.worldometers.info/coronavirus

Thailand, Turkey, and United Arab Emirates. The European countries include Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Moldova, Netherlands, Norway, Poland, Portugal, Romania, Russia, Serbia, Slovenia, Spain, Sweden, Switzerland, and Ukraine. The Africa countries include Algeria, Morocco, South Africa, and Tunisia. The countries in America include Argentina, Brazil, Canada, Chile, Colombia, Dominican Republic, Ecuador, Mexico, Panama, and Peru. In addition, we have the Oceania countries, Australia and New Zealand, in the sample. The data contain the number of deaths and the number of confirmed cases, together with 12 covariates displayed in Table 1. The data, available from the public sources listed in Table 1, were collected between February 27, 2020 and April 9, 2020.

All of the covariates are continuous variables. The covariate x_1 is defined to be the percentage of individuals age 65 or over in a country where the total population sizes counts all residents regardless of legal status or citizenship. The covariate x_2 is the prevalence of smoking which is the percentage of males and females age 15 and over who smoke any tobacco product. The covariate x_3 is the population weighted exposure to ambient PM2.5 pollution, defined as the average level of exposure of a nation's population to concentrations of suspended particles measuring less than 2.5 microns in aerodynamic diameter; exposure is calculated by weighting mean annual concentrations of PM2.5 by population in both urban and rural areas. The covariate x_4 is the age standardized death rate per $100,000$ individuals for both males and females. The covariate x_5 is the disease burden due to all chronic respiratory diseases, including silicosis, asthma, and lung disease. The covariate x_6 is the number of medical doctors, including general physicians and medical specialist per 1000 residents in a country. The covariate x_7 is the number of hospital beds, including inpatient beds available in public, private, general, and specialized hospitals and rehabilitation centers per 1000 residents in a country. The covariate x_8 is the mean systolic blood pressure (in mmHg) for adults 18 years and older, which is related to the level of the hypertension in a country. The covariates x_9 and x_{10} are the semi-annual average temperature and relative humidity from November of the previous year to April of its following year, respectively. This consideration is driven by the fact that the first case of the COVID-19 was discovered in early December of 2019 and the data we consider span to the end of April of 2020. The covariates x_{11} and x_{12} are the number of serious cases and the total number of tests per million people in a

Table 2: Summary of the data sets of Studies 1 and 2. Min and Max stand for minimum and maximum, respectively; and Mean represents the average.

Study	The Number of Deaths			The Number of Cases		
	Min	Mean	Max	Min	Mean	Max
Study 1	0	4	43	400	431	670
Study 2	0	38	291	200	1872	13531

country, respectively.

As the outbreak of COVID-19 starts at different times for different regions, there is a time lag to obtain the number of deaths and the number of confirmed cases. To circumvent this issue, we conduct a cross-sectional study by defining the time point as the day of the first 400 confirmed cases, and we call this "Study 1". For comparison, we take a second time point, defined as the 14 days after the first 100 cases are confirmed, which is partially driven by the fact that the maximum incubation time for COVID-19 is about 14 days (e.g., He et al., 2020), and we call this "Study 2". The minimum, mean, and maximum values of the number of deaths and the number of confirmed cases for the two studies are shown in Table 2.

3 Framework and Algorithm

3.1 Notation and Framework

Let $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ be the vector of covariates which follows a p-dimensional elliptical distribution with $\mathrm{E}(X) = 0$ and covariance matrix $\mathrm{Cov}(X) = \Sigma$, where the diagonal elements of Σ are $\mathrm{Var}(X_j) = 1$ for $j = 1, \ldots, p$. Let Y be the response variable. We consider the multiple index model for Y and X (Li, 1991):

$$Y = f\left(\beta_1^\top X, \beta_2^\top X, \ldots, \beta_d^\top X, \epsilon\right).$$

where ϵ is the error term independent of X with $E(\epsilon) = 0$ and $\{\beta_1, \ldots, \beta_d\}$ are unknown projection vectors with $\beta_k = (\beta_{k1}, \ldots, \beta_{kp})^\top$ for $k = 1, \ldots, d$. Here d is unknown but is regarded to be much smaller than p. Several methods are available in the literature for determining the structural dimension d, such as the asymptotic test (Li, 1991), the permutation test (Cook and Yin, 2001), the information criterion (Zhu et al., 2006), and the estimation algorithm (Lin et al., 2019). In our algorithm to be described in Section 3.2, we modify the estimation algorithm of Lin et al. (2019) by combing with the K-means method (MacQueen, 1967) to choose a suitable value for d.

Let $B = (\beta_1, \ldots, \beta_d)$ which is a nonidentifiable $p \times d$ matrix in the sense that the vectors $\beta_1, \ldots, \beta_d$ are not unique. However, the space spanned by the columns of B, called the central space and denoted $\mathrm{col}(B)$, can be uniquely determined. Li (1991) proposed the sliced inverse regression (SIR) procedure to estimate the central subspace $\mathrm{col}(B)$ without knowing function $f(\cdot)$. The SIR method is summarized as follows.

Assume that there exist n independent random pairs (y_i, x_i) whose distributions are identical to (Y, X), where $i = 1, \ldots, n$. We first divide them into H equal sized slices based on the order statistics $y_{(i)}$ for $i = 1, \ldots, n$, where H is a user-specified number of slices. Let $c = \lceil n/H \rceil$, we define $y_{h,l} = y_{(c(h-1)+l)}$ and $x_{h,l} = x_{(c(h-1)+l)}$, and then we rewrite the data as $(y_{h,l}, x_{h,l})$, where

h is the slice number and l is the order number of a sample in the hth slice, and $x_{(r)}$ is the covariate vector corresponding to $y_{(r)}$ for an index r. Let $\bar{x}_h = \frac{1}{c}\sum_{l=1}^{c} x_{h,l}$ be the sample mean in the hth slice, then $\Lambda = \mathrm{var}\{\mathrm{E}(X|Y)\}$ can be estimated by

$$\hat{\Lambda} = \frac{1}{H}\sum_{h=1}^{H} \bar{x}_h \bar{x}_h^\top = \frac{1}{H}\mathcal{X}_H \mathcal{X}_H^\top, \tag{1}$$

where $\mathcal{X}_H = (\bar{x}_1, \ldots, \bar{x}_H)$ is a $p \times H$ matrix. Let $\hat{V}$ be the matrix of the top d eigenvectors of $\hat{\Lambda}$ and let $\mathrm{col}(\hat{V})$ be the subspace formed by the column vectors of $\hat{V}$. Then the central subspace $\mathrm{col}(B)$ is estimated by $\hat{\Sigma}^{-1}\mathrm{col}(\hat{V})$ using the observed data, where $\hat{\Sigma}$ is the estimated covariance matrix of X.

3.2 Methodology and Algorithm

For ease of exposition, we assume $y_1 \le y_2 \le \cdots \le y_n$ for the sample $\{(y_i, x_i) : i = 1, \ldots, n\}$, where y_i is taken as the IFR of the ith country in our analysis of the COVID-19 data in Section 4. Let $M = I_H \otimes 1_c$ be the $n \times H$ matrix, where I_H is the $H \times H$ identity matrix, 1_c is the $c \times 1$ unit vector, and $\otimes$ represents the Kronecker product. Then we have $\mathcal{X}_H = xM/c$, where $x = (x_1^\top, \ldots, x_n^\top)$. Let $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$ be the d-top eigenvalues of $\hat{\Lambda}$ obtained by (1) and let $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_d)$ denote the corresponding eigenvectors of $\hat{\Lambda}$. Then by (1), we have

$$\hat{\lambda}_k \hat{\eta}_k = \frac{1}{H}\mathcal{X}_H \mathcal{X}_H^\top \hat{\eta}_k = \frac{1}{nc} xMM^\top x^\top \hat{\eta}_k$$

for $k = 1, \ldots, d$, and define a multivariate pseudo response $\tilde{y}$

$$\tilde{y} = \frac{1}{c} MM^\top x^\top \hat{\eta}\ \mathrm{diag}\left(\frac{1}{\hat{\lambda}_1}, \ldots, \frac{1}{\hat{\lambda}_d}\right).$$

To estimate the space spanned by $\{\beta_1, \ldots, \beta_d\}$, Lin et al. (2019) proposed the Lasso-SIR algorithm to find the sparse estimates, say, $\hat{\beta}_k$, for β_k with $k = 1, \ldots, d$. Specifically, for $k = 1, \ldots, d$, let

$$L_{\beta_k, k} = \frac{1}{2n}\|\tilde{y}_{*,k} - x^\top \beta_k\|_2^2 + \mu_k \|\beta_k\|_1, \tag{2}$$

where

$$\tilde{y}_{*,k} = \frac{1}{c\hat{\lambda}_k} MM^\top x^\top \hat{\eta}_k,$$

and μ_k is the tuning parameter. Then minimizing $L_{\beta_k,k}$ with respect to β_k gives the sparse estimates $\hat{\beta}_k$. Lin et al. (2019) showed that the Lasso penalized SIR regression (2) is asymptotically equivalent to the Lasso penalized linear regressions. The R package *glmnet* can be used to implement the Lasso penalized linear regression where the tuning parameters μ_k for $k = 1, \ldots, d$ are chosen by the cross-validation method.

The Lasso penalized linear regressions method does not have the selection consistency (Tibshirani, 1996; Zou, 2006). To overcome this limitation, Zou (2006) proposed the adaptive Lasso penalized linear regressions method which enjoys the selection consistency. The adaptive Lasso penalized methods can shrink the estimates of the parameters of unimportant covariates to 0 as the sample size $n \to \infty$, thus, removing all the unimportant covariates automatically and yielding sparse estimation results. Motivated by this, we propose the adaptive Lasso penalized SIR

regression. For $k = 1, \ldots, d$, we calculate the adaptive Lasso penalized SIR regression expression

$$\tilde{L}_{\beta_k,k} = \frac{1}{2n}\|\tilde{y}_{*,k} - x^\top \beta_k\|_2^2 + \mu_k \sum_{j=1}^{p} w_{kj}|\beta_{kj}|, \tag{3}$$

where the $w_{kj} = 1/|\tilde{\beta}_{kj}|^r$ with $r > 0$ are weights controlling the penalty for $k = 1, \ldots, d$ and $j = 1, \ldots, p$, the $\tilde{\beta}_{kj}$ are the estimates obtain from the unpenalized sliced inverse regression, and r can be simply set as 1. To obtain the sparse estimate $\hat{\beta}_k$, we minimize $\tilde{L}_{\beta_k,k}$ with respect to β_{kj}. Specifically, we carry out the following steps.

Step 1. Standardize the covariates x_j by subtracting the mean and dividing the standard deviation for $j = 1, \ldots, p$; calculate $\hat{\Lambda} = \frac{1}{n}\mathcal{X}_H\mathcal{X}_H^\top$ and find the top d eigenvalues and the corresponding eigenvectors, $\hat{\lambda}_k$ and $\hat{\eta}_k$, for $k = 1, \ldots, d$;

Step 2. For $k = 1, \ldots, d$, calculate

$$\tilde{y}_{*,k} = \frac{1}{c\hat{\lambda}_k} M M^\top x^\top \hat{\eta}_k;$$

Step 3. For $k = 1, \ldots, d$, solve the adaptive Lasso penalized SIR problem

$$\hat{\beta}_k = \arg\min_{\beta_{kj}} \left\{ \frac{1}{2n}\|\tilde{y}_{*,k} - x^\top \beta_k\|_2^2 + \mu_k \sum_{j=1}^{p} w_{kj}|\beta_{kj}| \right\}; \tag{4}$$

Step 4. Estimate the central subspace $\mathrm{col}(B)$ by $\mathrm{col}(\hat{B}) = \mathrm{col}(\hat{\beta}_1, \ldots, \hat{\beta}_d)$.

Steps 1 and 2 are easy to implement. For Step 3, we first fit the data using unpenalized sliced inverse regression to obtain $\tilde{\beta}_{kj}$ for $k = 1, \ldots, d$ and $j = 1, \ldots, p$. The Lasso penalized linear regression was well studies by Tibshirani (1996), Efron et al. (2004) and Friedman et al. (2010). Since we can convert the optimization problem (4) to the optimization problem of adaptive Lasso penalized linear regression, we solve the optimization problem (4) using the R package *glmnet* with the argument *alpha* = 1. The tuning parameters μ_k with $k = 1, \ldots, d$ are chosen by the cross-validation method. By solving the minimization problem in Step 3, we obtain a sparse estimate of β. Since $\mathrm{col}(\hat{B}) = \mathrm{col}(\hat{\beta}_1, \ldots, \hat{\beta}_d)$ estimates the central subspace $\mathrm{col}(B)$, the jth covariate x_j is considered to be unimportant if $\hat{\beta}_{kj} = 0$ for $k = 1, \ldots, d$.

4 Data Analysis

We analyze the COVID-19 data described in Section 2 using the proposed adaptive Lasso penalized SIR regression method, denoted ALSIR. In comparison, we also apply the linear regressions method and the unpenalized sliced inverse regression method, denoted LR and SIR, respectively. As discussed in Section 2, we analyze the two data sets of Studies 1 and 2 separately. To make the covariates unit-less and of the same scale for the penalty function, we standardize all the covariates by subtracting the mean and dividing the standard deviation before applying these methods.

4.1 Analysis Results with $H = 5$

In this subsection, we set $H = 5$ when implementing the ALSIR method to analyze the data. We report analysis results in Table 3 which includes the parameter estimates (Est.), the associated

Table 3: Analysis results for Study 1 and Study 2 using the linear regression (LR), unpenalized sliced inverse regression (SIR) and adaptive Lasso penalized SIR (ALSIR) methods. The entries for "Est.", "SE", and "ALSIR" are in percent.

Covariate	Study 1					Study 2				
	LR			SIR	ALSIR	LR			SIR	ALSIR
	Est.	SE	P-value			Est.	SE	P-value		
x_1	0.571	0.636	0.374	0.337	0.000	0.652	0.811	0.425	−0.377	0.496
x_2	−0.326	0.364	0.375	−0.146	0.000	−0.312	0.465	0.505	0.148	0.000
x_3	−0.106	0.411	0.797	−0.001	0.000	−0.131	0.524	0.803	0.188	0.000
x_4	0.980	0.419	0.023	0.687	0.349	0.860	0.534	0.113	−0.627	0.429
x_5	−0.015	0.400	0.970	−0.123	0.000	0.099	0.506	0.846	−0.049	0.000
x_6	−0.505	0.393	0.204	−0.178	−0.100	−0.387	0.500	0.443	0.208	−0.261
x_7	−0.376	0.407	0.359	0.003	0.000	−0.414	0.519	0.428	0.165	−0.372
x_8	−0.153	0.317	0.633	−0.212	0.000	−0.119	0.404	0.770	0.177	0.000
x_9	0.141	0.418	0.738	0.311	0.000	0.302	0.533	0.574	−0.425	0.371
x_{10}	−0.175	0.346	0.614	−0.111	0.000	0.067	0.441	0.880	−0.086	0.000
x_{11}	0.459	0.276	0.102	0.377	0.000	0.410	0.352	0.249	−0.218	0.360
x_{12}	−0.117	0.325	0.721	−0.226	−0.120	−0.352	0.415	0.400	0.260	−0.016

standard errors (SE), and the P-values if a linear relationship is assumed. By fitting the data of Studies 1 and 2 with a modified estimation algorithm of Lin et al. (2019), we estimate the structural dimension of the subspace to be $d = 1$. Therefore, we report only one direction vector $(\hat{\beta}_{11}, \ldots, \hat{\beta}_{1p})^\top$ as the estimated basis for the central subspace in the analysis; the jth covariate is considered to be unimportant if $\hat{\beta}_{1j}$ is shrunk to zero. That is, the jth feature does not affect the fatality of the COVID-19 if $\hat{\beta}_{1j} = 0$. Since the ALSIR method provides estimates of direction vectors for the basis of the central subspace derived from the multiple index model, positive or negative signs of estimates do not indicate positive or negative association of the covariate with the response.

The linear regressions method shows that except for the covariate x_4, none of the covariates are significantly associated with the IFR of COVID-19 for Study 1 data if we take the significant level to be 0.05. The unpenalized sliced inverse regression method does not remove unimportant risk factors automatically. On the contrary, the ALSIR method yields different results which implies that the linear relationship may not be appropriate. It shows that in Study 1, the covariates x_4, x_6 and x_{12} are important features for affecting the fatality. That is, the cardiovascular disease, the number of physicians, and the number of tests are closely associated with the fatality at the early stage of the outbreak. Clerkin et al. (2020) concluded that the patients with cardiovascular disease are at higher risk of fatality. Compared with the number of hospital beds, the number of doctors plays an important role in explaining the fatality rate. The fatality rate at the early stage of the COVID-19 does not seem to differ for the populations with different age structures.

Regarding the results for Study 2, our method shows that the covariates x_1, x_4, x_6, x_7, x_9, x_{11} and x_{12} are significant. That is, the age structure, cardiovascular disease, the number of physicians, the number of hospital beds, average temperature, the number of serious cases, and the number of tests are all important features for the fatality rate of the infected populations.

Onder et al. (2020) found that the populations with a higher proportion of older people have a higher fatality rate. The cardiovascular disease is an important feature for IFR of the infected populations. The number of medical doctors and the number of hospital beds per 1000 residents are all important for explaining the fatality rate. Studying the data in Wuhan city, China, Ma et al. (2020) suggested that temperature has strong positive correlation with the fatality of COVID-19, while some other research indicates that high temperature could prevent the spread of COVID-19 (e.g., Wang et al., 2020a,c). In our analysis here, the covariate x_9 is semi-annual average temperature collect from November to April, with the maximum, median and minimum values being 26.583°C, 6.717°C and −11.583°C respectively, and most countries in the data sets are in winter or spring during this time period. Our analysis indicates that temperature is associated with the fatality of COVID-19. Wang et al. (2020c) showed that the most suitable survival temperature for the coronvirus of COVID-19 is 8.72°C. The smoking prevalence and PM2.5 air pollution do not associate with the fatality rate of COVID-19. The analysis results support that the number of serious cases and the number of tests are important features for the fatality of COVID-19. The analysis results show that the humidity does not associate with the IFR of COVID-19. Chronic respiratory diseases and blood pressure (hypertension) do not have a strong association with the fatality.

In summary, the cardiovascular disease, the number of physicians and the number of tests are important features for both studies, while the smoking prevalence, PM2.5 air pollution, chronic respiratory disease, blood pressure, and humidity do not seem to be associated with the fatality. The age structure, the number of hospital beds, and the number of serious cases play different roles in explaining the fatality; they are important features for Study 2 but not for Study 1. For both studies, the chronic respiratory diseases and blood pressure (hypertension) are not considered as important for explaining the COVID-19 fatality.

4.2 Sensitivity Analysis

To compare the performance of the different methods, we further provide a "bootstrap inference" for the LR, SIR and ALSIR methods, as suggested by a referee. To be specific, we independently resample 1500 bootstrap samples from the initial data of Studies 1 and 2. We fit the 1500 bootstrap samples using these methods and obtain 1500 analysis results for each of the two studies accordingly. Then we calculate the resultant standard errors and report the results in Table 4. When using the proposed method, we choose $H = 5$ as in Section 4.1. Unsurprisingly, the results provided by the three methods are different. Interestingly, the proposed method seems to provide more stable results than the LR and SIR methods, because its associated standard errors are smaller than those produced by the LR and SIR methods.

As discussed by Liquet and Saracco (2012), the SIR method may be sensitive to the choice of the number H of slices. To see if the same issue is related to the ALSIR method, here we investigate the effect of the choice of H on the estimation results of the proposed method by considering three values of H: $H = 3$, 5 or 8, where the structural dimension d is taken as 1, as in Section 4.1. The parameter estimation results are recorded in Table 5, which clearly shows that the performance of the ALSIR is sensitive to the choice of H.

5 Discussion

In this paper, we propose an adaptive Lasso penalized sliced inverse regression method for the multiple index model to analyze the COVID-19 data. The proposed method is flexible in the

Table 4: Comparison of the performance of the linear regression (LR), unpenalized sliced inverse regression (SIR), and adaptive Lasso penalized SIR (ALSIR) methods using bootstrap resampling: The entries represent the empirical standard errors. The entries of LR and ALSIR are in percent.

Covariate	Study 1			Study 2		
	LR	SIR	ALSIR	LR	SIR	ALSIR
x_1	0.618	0.471	0.279	0.818	0.374	0.205
x_2	0.444	0.207	0.127	0.516	0.230	0.125
x_3	0.494	0.251	0.158	0.628	0.248	0.138
x_4	0.478	0.414	0.309	0.541	0.380	0.209
x_5	0.697	0.285	0.202	1.143	0.293	0.174
x_6	0.532	0.258	0.193	0.692	0.265	0.166
x_7	0.543	0.287	0.183	0.682	0.300	0.231
x_8	0.289	0.214	0.147	0.373	0.178	0.101
x_9	0.452	0.262	0.161	0.555	0.332	0.206
x_{10}	0.344	0.205	0.144	0.428	0.222	0.129
x_{11}	0.425	0.223	0.176	0.544	0.250	0.188
x_{12}	0.661	0.253	0.190	1.018	0.320	0.193

Table 5: Sensitivity of the performance of the ALSIR to the choice of H. The entries are parameter estimates in percent.

Covariate	Study 1			Study 2		
	$H = 3$	$H = 5$	$H = 8$	$H = 3$	$H = 5$	$H = 8$
x_1	0.089	0.000	−0.661	0.000	0.496	0.000
x_2	−0.099	0.000	0.132	0.000	0.000	0.000
x_3	0.000	0.000	0.031	0.000	0.000	0.000
x_4	0.225	0.349	0.000	0.058	0.429	−0.237
x_5	0.000	0.000	0.000	0.000	0.000	0.000
x_6	−0.084	−0.100	0.000	0.000	−0.261	0.313
x_7	0.079	0.000	0.000	0.000	−0.372	0.000
x_8	−0.074	0.000	0.000	0.000	0.000	0.000
x_9	0.111	0.000	0.000	0.073	0.371	0.000
x_{10}	0.023	0.000	−0.270	0.000	0.000	0.000
x_{11}	0.095	0.000	0.000	0.022	0.360	−0.230
x_{12}	−0.045	−0.120	−0.128	−0.066	−0.016	0.635

Table 6: Pairwise Pearson correlation coefficients between the covariates.

Covariate	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
x_1	0.444	−0.634	−0.233	−0.536	0.691	0.643	0.022	−0.618	0.367	0.229	0.015
x_2	-	−0.046	0.344	−0.115	0.364	0.502	0.237	−0.457	0.133	0.100	−0.012
x_3		-	0.209	0.528	−0.463	−0.324	0.071	0.338	−0.514	−0.097	−0.031
x_4			-	−0.050	−0.107	0.229	0.454	−0.227	0.085	−0.255	−0.212
x_5				-	−0.312	−0.424	−0.236	0.420	−0.408	−0.085	0.378
x_6					-	0.503	0.079	−0.566	0.143	0.164	0.197
x_7						-	0.223	−0.588	0.127	0.056	−0.073
x_8							-	−0.137	0.222	−0.237	−0.170
x_9								-	−0.341	−0.081	−0.144
x_{10}									-	0.021	0.007
x_{11}										-	−0.046

sense that there is no need to specify the model structure of $f(\cdot)$ function for the multiple index model, which protects us from potential model misspecification. We apply the proposed method with H set as 5 to identify risk factors associated with the COVID-19 fatality, which shows that the cardiovascular disease, the number of doctors and the number of tests are important for both cross-sectional studies. However, as demonstrated in Section 4.2, the performance of proposed ALSIR method is sensitive to the choice of H, a phenomenon that also occurs with the SIR method (Liquet and Saracco, 2012). When interpreting the analysis results, one must be aware of this uncertainty and regard the inference procedure as a kind of *conditional* analysis in the sense that H is set as a given value.

Our cross-sectional study examines the data sets at two time points of the outbreak which are defined as the time of first 400 confirmed cases and the 14 days after the first 100 cases are confirmed. This consideration mainly focuses on studying the early stage of the pandemic with the incubation period taken into account. One may, of course, consider other time points and repeat the same investigation, and different findings may be uncovered as data become richer. More generally, it is interesting to extend this development to the longitudinal data framework and explore the dynamic relationship between the fatality and risk factors.

Several important issues should be acknowledged, as noticed by the referees and discussed by other authors (e.g., He et al., 2020). Due to asymptomatic infections and limited test capacity, the reported numbers of cases with COVID-19 are typically smaller than the true numbers of infections for various countries. Additionally, other aspects such as varying incubation periods make the data error-prone. To help understand the underlying truth, it is useful to conduct sensitivity analyses by modifying the development here to accommodate the feature of error-contaminated data (Carroll et al., 2006; Yi, 2017).

In the analysis, one concern pertains to the collinearity of covariates. In the data we analyze, the correlation for paired covariates are fairly small, as suggested by the values in Table 6. Other concerns may be related to the inclusion of potential risk factors for the analysis. For example, one may ask: have we exhausted all possible risk factors for the fatality? Do we need to worry about confounding issues among the risk factors? Questions like these warrant careful studies. As the outbreak continues to exacerbate, more data become available and in-depth studies are possible to help us better understand the characteristics of the COVID-19.

Supplementary Materials

The data and R code needed to reproduce the results in this paper can be found at the *Journal of Data Science* website.

Acknowledgments

The authors thank the editor and the referees for their comments on the initial submission. This research is partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) as well as the Rapid Response Program – COVID-19 of the Canadian Statistical Sciences Institute (CANSSI). Yi is Canada Research Chair in Data Science (Tier 1). Her research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.

References

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC Press, Boca Raton, Florida.

Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *The Lancet*, 395(10223): 507–513.

Clerkin KJ, Fried JA, Raikhelkar J, Sayer G, Griffin JM, Masoumi A, et al. (2020). COVID-19 and cardiovascular disease. *Circulation*, 141(20): 1648–1655.

Cook RD (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. John Wiley & Sons, New York.

Cook RD, Li B, et al. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2): 455–474.

Cook RD, Weisberg S (1991). Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association*, 86(414): 328–332.

Cook RD, Yin X (2001). Dimension-reduction and visualization in discriminant analysis. *Australian & New Zealand Journal of Statistics*, 43(2): 147–199.

Efron B, Hastie T, Johnstone I, Tibshirani R (2004). Least angle regression. *The Annals of Statistics*, 32(2): 407–499.

Epidemiology Working Group for NCIP Epidemic Response (2020). The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. *Chinese Journal of Epidemiology*, 41(2): 145–151.

Friedman J, Hastie T, Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1–22.

Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. (2020). Clinical characteristics of 2019 novel coronavirus infection in China. MedRxiv preprint: https://doi.org/10.1101/2020.02.06.20020974.

He W, Yi GY, Zhu Y (2020). Estimation of the basic reproduction number, average incubation time, asymptomatic infection rate, and case fatality rate for COVID-19: Meta-analysis and sensitivity analysis. *Journal of Medical Virology*. Forthcoming, https://doi.org/10.1002/jmv.26041.

Hu Z, Ge Q, Jin L, Xiong M (2020). Artificial intelligence forecasting of COVID-19 in China. ArXiv preprint: https://arxiv.org/abs/2002.07112.

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223): 497–506.

Ji Y, Ma Z, Peppelenbosch MP, Pan Q (2020). Potential association between COVID-19 mortality and health-care resource availability. *The Lancet Global Health*, 8(4): e480.

Li KC (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414): 316–327.

Li KC (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420): 1025–1039.

Li L (2007). Sparse sufficient dimension reduction. *Biometrika*, 94(3): 603–613.

Li L, Nachtsheim CJ (2006). Sparse sliced inverse regression. *Technometrics*, 48(4): 503–510.

Lin Q, Zhao Z, Liu JS (2019). Sparse sliced inverse regression via Lasso. *Journal of the American Statistical Association*, 114(528): 1726–1739.

Liquet B, Saracco J (2012). A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches. *Computational Statistics*, 27(1): 103–125.

Ma Y, Zhao Y, Liu J, He X, Wang B, Fu S, et al. (2020). Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. *Science of the Total Environment*, 724(1): 138226.

MacQueen J (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 281–297. Oakland, CA, USA.

Ni L, Cook RD, Tsai CL (2005). A note on shrinkage sliced inverse regression. *Biometrika*, 92(1): 242–247.

Onder G, Rezza G, Brusaferro S (2020). Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *Journal of the American Medical Association*, 323(18): 1775–1776.

Shi S, Qin M, Shen B, Cai Y, Liu T, Yang F, et al. (2020). Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan, China. *JAMA Cardiology*. Forthcoming, http://doi.org/10.1001/jamacardio.2020.0950.

Tibshirani R (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1): 267–288.

Wang J, Tang K, Feng K, Lv W (2020a). High temperature and high humidity reduce the transmission of COVID-19. SSRN preprint: https://dx.doi.org/10.2139/ssrn.3551767.

Wang L, Zhou Y, He J, Zhu B, Wang F, Tang L, et al. (2020b). An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China (with discussion). *Journal of Data Science*, 18(3): 409–454.

Wang M, Jiang A, Gong L, Luo L, Guo W, Li C, et al. (2020c). Temperature significant change COVID-19 transmission in 429 cities. MedRxiv preprint: https://doi.org/10.1101/2020.02.22.20025791.

Wang Z, Yang B, Li Q, Wen L, Zhang R (2020d). Clinical features of 69 cases with coronavirus disease 2019 in Wuhan, China. *Clinical Infectious Diseases*. Forthcoming, https://doi.org/10.1093/cid/ciaa272.

Wu Y, Li L (2011). Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Statistica Sinica*, 21(3): 707–730.

Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C, et al. (2020). Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *The Lancet Respiratory Medicine*, 8(4): 420–422.

Yi GY (2017). *Statistical Analysis with Measurement Error or Misclassification*. Springer, New York.

Zhang C, Shi L, Wang FS (2020). Liver injury in COVID-19: Management and challenges. *The Lancet Gastroenterology & Hepatology*, 5(5): 428–430.

Zhou J, He X, et al. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *The Annals of Statistics*, 36(4): 1649–1668.

Zhu L, Miao B, Peng H (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474): 630–643.

Zou H (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476): 1418–1429.