Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 295-306

A Mixture of Regressions Model of COVID-19 Death Rates and Population Comorbidities

M. Maleki[1], G. J. McLachlan[2], R. Gurewitsch[3], M. Aruru[4,5], S. Pyne[3,5,6]

[1]*Faculty of Mathematics and Statistics, Department of Statistics, University of Isfahan, Isfahan, Iran*
[2]*Department of Mathematics, University of Queensland, St. Lucia, Brisbane, Australia*
[3]*Public Health Dynamics Laboratory, University of Pittsburgh, Pittsburgh, USA*
[4]*Program Evaluation and Research Unit, University of Pittsburgh School of Pharmacy, Pittsburgh, USA*
[5]*Health Analytics Network, PA, USA*
[6]*Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, USA*

Received: 08 July 2020; Revised: 22 July 2020; Accepted: 24 July 2020

Abstract

As the COVID-19 pandemic spread worldwide, it has become clearer that prevalence of certain comorbidities in a given population could make it more vulnerable to serious outcomes of that disease, including fatality. Indeed, it might be insightful from a health policy perspective to identify clusters of populations in terms of the associations between their prevalent comorbidities and the observed COVID-19 specific death rates. In this study, we described a mixture of polynomial time series (MoPTS) model to simultaneously identify (a) three clusters of 86 U.S. cities in terms of their dynamic death rates, and (b) the different associations of those rates with 5 key comorbidities among the populations in the clusters. We also described an EM algorithm for efficient maximum likelihood estimation of the model parameters.

Keywords: Mixture of regressions; EM; Death rate; Comorbidities; COVID-19.

1. Introduction

COVID-19 is an acute, respiratory disease due to novel coronavirus SARS-CoV-2, and similar to previous diseases such as Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS), it can lead to respiratory failure and death (CDCa, 2019). In the absence of any proven medical treatment for COVID-19, and in the face of acute shortage of critical care capacity, including ventilators, during peaks of incidence, as was witnessed in some countries such as Italy, it is of critical importance for any local administration to evaluate and stratify the risk levels in terms of its population comorbidities.

As per current understanding (CDCb, 2020), individuals at-risk for severe illness from COVID-19 include people who are or have: (a) 65 years and older, (b) living in nursing home or long-term care facilities, (c) chronic lung disease or moderate to severe asthma, (d) serious

Corresponding author: Saumyadipta Pyne
Email: spyne@pitt.edu

heart conditions, (e) immuno-compromised, (f) severe obesity, (g) diabetes, (h) chronic kidney disease undergoing dialysis, (i) hemoglobin disorders, and (j) liver disease.

Early retrospective studies from China, Italy, and the U.S. revealed the prevalence of comorbidities in exacerbation of disease resulting in poor outcomes. Retrospective analysis of 1,590 confirmed COVID-19 cases between December 11, 2019 and January 31, 2020 across 31 provinces/autonomous regions in China indicated that 25.1% of the cases had at least one comorbidity. Specific prevalence included hypertension (16.9%), other cardiovascular diseases (53.7%) and diabetes (8.2%), whereas asthma, COPD, kidney diseases, immunodeficiency were below 1% (Guan *et al*., 2020). In a larger study across China, 72,314 patient records until February 11, 2020 were examined of which 44,672 were confirmed cases and further evaluated for comorbidities. (Feng *et al*., 2020) More than 50% of the records were missing data on comorbidities. Of available data on confirmed cases, comorbidities included hypertension (12.8%), diabetes (5.3%), and less than 4% for cardiovascular disorders, chronic respiratory disease, and cancers.

The case-fatality ratio (CFR) is a useful indicator of survival prognosis among critically ill patients. During the peak of COVID-19 infections in China, the highest CFR was 14.8% among patients over age 80 years. CFR for patients with comorbid conditions was 10.5% for those with cardiovascular disease, 7.3% for diabetes, 6.3% for chronic respiratory disease, 6% for hypertension and 5.6% for cancer, when compared to 0.9% for patients with no comorbid conditions. CFR for critical cases was 49% indicting the need to mitigate risks earlier in the infection. A meta-analysis of 6 COVID-19 studies indicated hypertension, diabetes, COPD, cardiovascular disease, and cerebrovascular disease as major risk factors for COVID-19 patients (Wang *et al*., 2020). Another meta-analysis of pooled data from 8 studies in China, indicated a higher prevalence of hypertension and diabetes among patients with severe cases of COVID-19 disease (Yang *et al*., 2020).

As the pandemic spread, Italy was next to experience large caseloads and CFR. COVID-19 cases peaked in March to 3000-6500 new cases and 350-900 deaths daily. Comorbidities observed in deceased patients as of April 29, 2020 included hypertension (69.2%), diabetes (31.8%), cardiovascular conditions (22%), COPD (16.9%), obesity (11.6%) among others (Statista, 2020). In China and Italy, hypertension, diabetes, cardiovascular conditions, chronic respiratory disease accounted for the top comorbidities associated with severe outcomes (Chen et al., 2020).

The first COVID-19 case in the U.S. was reported from Washington State on January 31, 2020. Since then, the virus spread to the east coast and the rate of infections in New York exceeded every other state. As of April 30, 2020, New York had more than 30% of all the U.S. cases. A retrospective analysis of 5,700 records of patients hospitalized with COVID-19 in the New York City area revealed that while 6.1% of the patients had no comorbidities, and 6.3% had 1 comorbidity, as high as 88% had more than one comorbidity. This study further indicated that older patients, men, and those with pre-existing hypertension and/or diabetes were highly prevalent among those hospitalized for COVID-19 (Richardson et al., 2020). This pattern was similar to observations from China and Italy. In yet another observational cohort study at two New York hospitals, as of April 28, 2020, 39% of patients who were critically ill with COVID-19 had died, and similar to studies in China and Italy, hypertension and cardiopulmonary comorbidities were found to be associated with increased mortality (Cummings *et al*., 2020).

In this study, we modeled COVID-19 death rates and key comorbidities for $N = 86$ U.S. cities. We used a mixture of regressions modeling approach to simultaneously identify the

clusters of cities in terms of their dynamic death rates, which had different associations with the prevalence of 5 key comorbidities among the populations in those clusters. We provided an EM algorithm for efficient maximum likelihood estimation of the model parameters. While the use of mixture of regressions models is well-known, [*e.g.*, Jones and McLachlan (1992) and references in McLachlan and Peel (2000)], we think that the combination of finite mixtures, linear regression and polynomial modeling of time series makes our approach ideally suited for the present problem. We described our mixture of polynomial time series (MoPTS) model and the datasets in the next section, followed by the modeling results and discussion.

2. Materials and Methods

2a. Datasets

Comorbidities Data: The Centers for Disease Control and Prevention (CDC) has conducted each year, since 1984, the Behavioral Risk Factor Surveillance System (BRFSS), a cross-sectional representative telephone survey of U.S. adults (18 years or above) regarding their health conditions and behaviors. BRFSS now collects data from all 50 U.S. states and the districts of Columbia, Guam, and Puerto Rico. Both landline and cellular phone numbers are sampled in the survey and aggregated results are presented each year (CDCc, 2020). Health risk questions range from infectious diseases to chronic diseases and behaviors. The fixed core of the survey consists of a standard set of questions used by all states to include demographics and current health behaviors, *e.g.* tobacco use. In this study, we focus on $p = 5$ comorbidities collected by BRFSS and reported at the county level: diabetes, obesity, coronary heart disease (CHD), hypertension, and chronic obstructive pulmonary disease (COPD).

COVID-19 Data: Based on cumulative COVID-19 deaths data from the Johns Hopkins Coronavirus Resource Center (COVID-19 U.S. cases, 2020), we compiled time series data on daily deaths from the disease for the U.S. counties by their 5-digit FIPS code or county name. Since a single county may consist of multiple cities, we include the list of all city labels within each aggregate group to represent a greater metropolitan area. While we denote such metropolitan areas as "cities" for convenience, we identify them by their underlying county FIPS codes. In this study, we used the data for $N = 86$ cities across the U.S. which had at least 100 reported deaths from COVID-19 by May 31, 2020. We excluded New York City from our clustering as it is an outlier in terms of uniquely high incidence and extreme population heterogeneity, which could skew the model with its own singleton cluster. For each included city j $(j = 1, \dots, N)$, the daily number of deaths is given as a time series of $n_j = 93$ time points, over the period starting from 29 February, 2020, up to 31 May, 2020. For compatibility, the COVID-19-specific death rate was standardized as the number of deaths per 100,000 people using population data of the cities from the U.S. Census Bureau (U.S. Census Bureau).

2b. Mixture of Regressions Model

In this study, we used a mixture of polynomial time series (MoPTS) model to cluster the above-mentioned $(N = 86)$ cities in terms of their dynamic death rates into an optimal number (g) of clusters (identified by the mixture model components $C_1, \dots C_g$), and investigate the differences among the estimated regression coefficients of 5 known comorbidities (given as 5 static city-specific covariates) across those clusters.

Let Y_{jt} be the random variable representing death rate at time t $(t = 1, \dots n_j)$ for city j $(j = 1, \dots, N)$. It is assumed that the mean $\mu_i(t)$ of each time series Y_{jt} in the i^{th} cluster C_i follows

$Y_{jt} - \mu_i(t) \overset{iid}{\sim} N(0, \sigma_i^2)$ in C_i $(i = 1, \dots, g)$ for all values of $t = 1, \dots n_j$ and $j = 1, \dots, N$. Further, let $\boldsymbol{x}_j$ be a p-dimensional vector that describes static values of p comorbidities for the city j $(j = 1, \dots, N)$. Also, let Z_j is a component-indicator random variable to determine that city j belongs to the component (or cluster) C_i with probability $Pr(Z_j = i) = \pi_i$ such that $\pi_i > 0$ and $\sum_{i=1}^{g} \pi_i = 1$ for $i = 1, \dots, g$ and $j = 1, \dots, N$. Thus, our mixture of regressions on time series modeled with order D polynomial and covariates has the form:

$$Y_{jt} = \boldsymbol{\alpha}_i^\top \boldsymbol{x}_j + \sum_{d=1}^{D} \beta_{id} t^d + \varepsilon_{jt}; \;\; \varepsilon_{jt} \sim N(0, \sigma_i^2) \text{ with probability } \pi_i; i = 1, \dots, g \qquad (1)$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{ip})^\top$ are regression coefficients corresponding to the p static covariates, $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \dots, \beta_{iD})^\top$ are coefficients corresponding to the polynomial time series.

Let $\boldsymbol{Y}_j = \left(Y_{j0}, Y_{j1}, \dots, Y_{jn_j}\right)^\top$ arise from a g-component mixture of order D polynomial time series with covariates model, denoted by MoPTS(D, g), if it can be characterized by the conditional density function

$$f(y_{jt} | \boldsymbol{x}_j, Z_j = i, \boldsymbol{\theta}_i) = \phi(y_{jt}; \boldsymbol{\alpha}_i^\top \boldsymbol{x}_j + \boldsymbol{\beta}_i^\top \boldsymbol{t}_D, \sigma_i^2), \;\; j = 1, \dots, N; t = 1, \dots, n_j, \qquad (2)$$

where $\boldsymbol{t}_D = (t, t^2, \dots, t^D)^\top$ and $\boldsymbol{\theta}_i = (\boldsymbol{\alpha}_i^\top, \boldsymbol{\beta}_i^\top, \sigma_i^2)^\top$.

Under the characterization (2), we can further characterize $\boldsymbol{Y}_j$ via the joint density function

$$f(\boldsymbol{y}_j | \boldsymbol{x}_j, Z_j = i, \boldsymbol{\theta}_i) = \sum_{i=1}^{g} \pi_i \prod_{t=1}^{n_j} \phi(y_{jt}; \boldsymbol{\alpha}_i^\top \boldsymbol{x}_j + \boldsymbol{\beta}_i^\top \boldsymbol{t}_D, \sigma_i^2), \;\; j = 1, \dots, N. \qquad (3)$$

Using the characterization (3), we can write the log-likelihood of an IID sample $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_N$ as

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{N} \log \left(\sum_{i=1}^{g} \pi_i \prod_{t=1}^{n_j} \phi(y_{jt}; \boldsymbol{\alpha}_i^\top \boldsymbol{x}_j + \boldsymbol{\beta}_i^\top \boldsymbol{t}_D, \sigma_i^2) \right), \qquad (4)$$

where the parameter $\boldsymbol{\theta} = \left(\pi_1, \dots, \pi_{g-1}, \boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_g^\top\right)^\top$ is traditionally estimated by maximization of $\ell(\boldsymbol{\theta})$ given by (4). However, in the absence of an analytical solution, we used an EM–type algorithm (Dempster *et al*., 1977; McLachlan and Peel 2000). Using log-likelihood criteria, we chose $D = 3$ and $g = 3$, *i.e*., our final fit produced a 3-component normal mixture regression model with the dependence of the mean on time modeled by a cubic polynomial.

2c. The observed information matrix

In this section, the observed (Fisher) information matrix of MoPTS, defined as $\mathbf{J}(\boldsymbol{\theta}|\boldsymbol{y}) = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$, where $\ell(\boldsymbol{\theta}) = \sum_{j=1}^{N} \ell_j(\boldsymbol{\theta})$, $\ell_j(\boldsymbol{\Theta}) = \log\left(\sum_{i=1}^{g} \pi_i \prod_{t=1}^{n_j} \phi(y_{jt}; \boldsymbol{\alpha}_i^\top \boldsymbol{x}_j + \boldsymbol{\beta}_i^\top \boldsymbol{t}_D, \sigma_i^2)\right)$, is obtained.

It is well known that, under mild regularity conditions, the covariance matrix of the ML estimates $\widehat{\boldsymbol{\Theta}}$ can be approximated by the inverse of $\mathbf{J}(\boldsymbol{\Theta}|\boldsymbol{y})$. Thus, following Basford *et al*. (1997), we approximated $\mathbf{J}(\boldsymbol{\Theta}|\boldsymbol{y})$ as

$$\mathbf{J}(\boldsymbol{\theta}|\boldsymbol{y}) \approx \sum_{j=1}^{N} \hat{\boldsymbol{s}}_j \hat{\boldsymbol{s}}_j^{\top}, \tag{5}$$

where $\hat{\boldsymbol{s}}_j = \frac{\partial \ell_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$, and consider now the vector $\hat{\boldsymbol{s}}_j$ which is partitioned into components corresponding to all the parameters in $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{s}}_j = \left(\hat{s}_{j,\pi_1}, \ldots, \hat{s}_{j,\pi_{g-1}}, \hat{\boldsymbol{s}}_{j,\boldsymbol{\alpha}_1}^{\top}, \ldots, \hat{\boldsymbol{s}}_{j,\boldsymbol{\alpha}_g}^{\top}, \hat{\boldsymbol{s}}_{j,\boldsymbol{\beta}_1}^{\top}, \ldots, \hat{\boldsymbol{s}}_{j,\boldsymbol{\beta}_g}^{\top}, \hat{s}_{j,\sigma_1^2}, \ldots, \hat{s}_{j,\sigma_g^2}\right)^{\top},$$

where its coordinate elements for $i = 1, \ldots, g$ are given by

$$\hat{s}_{j,\pi_i} = \frac{\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i) - \phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_g)}{f(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta})}, \quad \hat{\boldsymbol{s}}_{j,\boldsymbol{\Delta}_i} = \frac{\pi_i D_{\boldsymbol{\Delta}_i}\left(\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)\right)}{f(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta})},$$

where $f(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i \phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)$ and $\boldsymbol{\Delta}_i = \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$ and σ_i^2, for which $\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i) = \sum_{i=1}^{g} \pi_i \prod_{t=1}^{n_j} \phi(y_{jt}; \boldsymbol{\alpha}_i^{\top}\boldsymbol{x}_j + \boldsymbol{\beta}_i^{\top}\boldsymbol{t}_D, \sigma_i^2)$ and $D_{\boldsymbol{\Delta}_i}\left[\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)\right] = \partial\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)/\partial\boldsymbol{\Delta}_i$. Thus,

$$D_{\boldsymbol{\alpha}_i}\left[\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)\right] = \left(\frac{1}{\sqrt{2\pi}\sigma_i}\right)^{n_j} s_{xy} \exp\left(-\frac{1}{2}s_{yy}\right);$$

$$D_{\boldsymbol{\beta}_i}\left[\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)\right] = \left(\frac{1}{\sqrt{2\pi}\sigma_i}\right)^{n_j} s_{ty} \exp\left(-\frac{1}{2}s_{yy}\right);$$

$$D_{\sigma_i^2}\left[\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)\right] = \left(\frac{1}{\sqrt{2\pi}\sigma_i}\right)^{n_j} \frac{1}{2}\sigma_i^{2n_j-2}\left\{s_{yy} - n_j \exp\left(-\frac{1}{2}s_{yy}\right)\right\};$$

where $s_{yy} = \frac{1}{\sigma_i^2}\sum_{t=1}^{n_j}(y_{jt} - \boldsymbol{\alpha}_i^{\top}\boldsymbol{x}_j - \boldsymbol{\beta}_i^{\top}\boldsymbol{t}_D)^2$, $s_{xy} = \frac{1}{\sigma_i^2}\sum_{t=1}^{n_j}\boldsymbol{x}_j(y_{jt} - \boldsymbol{\alpha}_i^{\top}\boldsymbol{x}_j - \boldsymbol{\beta}_i^{\top}\boldsymbol{t}_D)$ and $s_{ty} = \frac{1}{\sigma_i^2}\sum_{t=1}^{n_j}\boldsymbol{t}_D(y_{jt} - \boldsymbol{\alpha}_i^{\top}\boldsymbol{x}_j - \boldsymbol{\beta}_i^{\top}\boldsymbol{t}_D)$.

2d. ML estimation

In this section, we develop an efficient EM algorithm for maximum likelihood (ML) estimation of the parameters of the MoPTS model, using an incomplete-data framework. To compute this procedure, we use the following hierarchical representation:

$$Y_{jt}|\boldsymbol{x}_j, Z_{ji} = 1 \overset{iid}{\sim} N(\boldsymbol{\alpha}_i^{\top}\boldsymbol{x}_j + \boldsymbol{\beta}_i^{\top}\boldsymbol{t}_D, \sigma_i^2), \quad i = 1, \ldots, g.$$

$$Pr(Z_{jt} = i) = \pi_i \ (i = 1, \ldots, g; \ j = 1, \ldots, N; t = 1, \ldots, n_j), \tag{6}$$

Let $\boldsymbol{y} = (\boldsymbol{y}_1^\top, , \dots, \boldsymbol{y}_N^\top)^\top$, $\boldsymbol{x} = (\boldsymbol{x}_1^\top, , \dots, \boldsymbol{x}_N^\top)^\top$ and $\boldsymbol{z} = (\boldsymbol{z}_1^\top, \dots, \boldsymbol{z}_N^\top)^\top$ for which $\boldsymbol{z}_j = (z_{j1}, \dots, z_{jg})^\top$; $j = 1, \dots, N$, so considering the complete data $\boldsymbol{y}_c = (\boldsymbol{y}^\top, \boldsymbol{x}^\top, \boldsymbol{z}^\top)^\top$ and using the hierarchical representation in (6) of the MoPTS(D, g) model, the complete data likelihood, ignoring the constant term, is given by

$$\begin{aligned}\mathcal{P}_c(\boldsymbol{\theta}|\boldsymbol{y}_c) &= \sum_{j=1}^{N}\sum_{i=1}^{g}\sum_{t=1}^{n_j} I_{[i]}(z_{jt})\{\log\pi_i + \log f(y_{jt}|\boldsymbol{x}_j, Z_j = i, \boldsymbol{\theta}_i)\} \\ &= \sum_{i=1}^{g}\log\pi_i \sum_{j=1}^{N}\sum_{t=1}^{n_j} I_{[i]}(z_{jt}) - \sum_{i=1}^{g}\log\sigma_i \sum_{j=1}^{N}\sum_{t=1}^{n_j} I_{[i]}(z_{jt}) \\ &\quad -\frac{1}{2}\sum_{i=1}^{g}\frac{1}{\sigma_i^2}\sum_{j=1}^{N}\sum_{t=1}^{n_j} I_{[i]}(z_{jt})(y_{jt} - \boldsymbol{\alpha}_i^\top\boldsymbol{x}_j - \boldsymbol{\beta}_i^\top\boldsymbol{t}_D)^2 \end{aligned} \quad (7)$$

where $I_{[i]}(z_{jt})$ is 1 if y_{jt} belongs to the ith component and zero otherwise, and $f(\cdot\,|\boldsymbol{x}_j, Z_j = i, \boldsymbol{\theta}_i)$ was defined in (2).

Starting from some initial value $\boldsymbol{\theta}^{(0)}$, the conditional expectation of (7), given the observed data, was computed using $\boldsymbol{\theta}^{(k)}$ for $\boldsymbol{\theta}$, can be written as

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= \sum_{i=1}^{g}\log\pi_i \sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\boldsymbol{\theta}^{(k)}) - \sum_{i=1}^{g}\log\sigma_i \sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\boldsymbol{\theta}^{(k)}) \\ &\quad -\frac{1}{2}\sum_{i=1}^{g}\frac{1}{\sigma_i^2}\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\boldsymbol{\theta}^{(k)})(y_{jt} - \boldsymbol{\alpha}_i^\top\boldsymbol{x}_j - \boldsymbol{\beta}_i^\top\boldsymbol{t}_D)^2, \end{aligned} \quad (8)$$

where

$$\gamma_{jit}(\boldsymbol{\theta}) = \frac{\pi_i\phi(y_{jt}; \boldsymbol{\alpha}_i^\top\boldsymbol{x}_j + \boldsymbol{\beta}_i^\top\boldsymbol{t}_D, \sigma_i^2)}{\sum_{s=1}^{g}\pi_i\phi(y_{jt}; \boldsymbol{\alpha}_s^\top\boldsymbol{x}_j + \boldsymbol{\beta}_s^\top\boldsymbol{t}_D, \sigma_s^2)}. \quad (9)$$

The posterior probability is the conditional probability that y_{jt} belongs to the i^{th} component given $\boldsymbol{x}_j$ and $\boldsymbol{y}_{j(t)}$ for $i = 1, \dots, g$; $j = 1, \dots, N$ and $t = 1, \dots, n_j$. To perform the M-step, under the restriction $\sum_{i=1}^{g}\pi_i = 1$, by constructing the Lagrangian $\Lambda(\boldsymbol{\theta}, \lambda) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) + \lambda(\sum_{i=1}^{g}\pi_i - 1)$, we maximize (8) in the EM algorithm by solving the equation corresponding to the first-order condition $\nabla\Lambda(\boldsymbol{\theta}, \lambda) = \boldsymbol{0}$, where ∇ is the gradient operator, which yields the following updates

$$\hat{\pi}_i^{(k+1)} = \frac{\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\hat{\boldsymbol{\theta}}^{(k)})}{N}, \quad (10)$$

$$\hat{\boldsymbol{\alpha}}_i^{(k+1)} = \left[\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\hat{\boldsymbol{\theta}}^{(k)})\boldsymbol{x}_j\boldsymbol{x}_j^\top\right]^{-1}\left[\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\hat{\boldsymbol{\theta}}^{(k)})\boldsymbol{x}_j\left(y_{jt} - \boldsymbol{t}_D^\top\hat{\boldsymbol{\beta}}_i^{(k)}\right)\right], \quad (11)$$

$$\hat{\boldsymbol{\beta}}_i^{(k+1)} = \left[\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\hat{\boldsymbol{\theta}}^{(k)})\boldsymbol{t}_D\boldsymbol{t}_D^\top\right]^{-1}\left[\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\hat{\boldsymbol{\theta}}^{(k)})\boldsymbol{t}_D\left(y_{jt} - \boldsymbol{x}_j^\top\hat{\boldsymbol{\alpha}}_i^{(k+1)}\right)\right], \quad (12)$$

$$\hat{\sigma}_i^{2(k+1)} = \frac{\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}\left(\hat{\boldsymbol{\theta}}^{(k)}\right)\left(y_{jt}-\boldsymbol{x}_j^{\top}\hat{\boldsymbol{\alpha}}_i^{(k+1)}-\boldsymbol{t}_D^{\top}\hat{\boldsymbol{\beta}}_i^{(k+1)}\right)^2}{\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}\left(\hat{\boldsymbol{\theta}}^{(k)}\right)}. \tag{13}$$

The steps of the above algorithm are iterated until a suitable convergence rule is satisfied, e.g., $\left|\ell\left(\hat{\boldsymbol{\theta}}^{(k+1)}\right)/\ell\left(\hat{\boldsymbol{\theta}}^{(k)}\right)-1\right| \leq \varepsilon$ for a pre-decided tolerance ε.

3. Results and Discussion

Based on the log-likelihood model selection criteria, our optimal 3-component MoPTS model with the dependence of the mean death rate on time was modeled by a cubic polynomial. Thus, we identified 3 clusters of U.S. cities specified by the different associations between their death rates and comorbidities. Figure 1 depicts the time series of each cluster along with the fitted cubic polynomial models. The differences in the death rates (y-axes) across the 3 clusters can be noted. Table 1 lists the counties and states corresponding to these 86 cities, and their cluster memberships. Table 2 shows the parameter coefficients of the MoPTS $(D = 3, g = 3)$ model as estimated by the EM algorithm, along with the corresponding standard deviations.

Cluster 1 contains 21 cities concentrated mostly in the northeastern part of the U.S. (Figure 2a) with death rates associated with diabetes (α_{11}) and hypertension (α_{14}). Cluster 2 consists of 29 cities (Figure 2b) that are more spatially distributed than those in cluster 1, and has relatively weaker association with CHD (α_{23}), diabetes (α_{21}) and COPD (α_{25}). Cluster 3 (Figure 2c), comprised of 36 cities, is the most geographically heterogeneous, which possibly explains why its association with the comorbidities hypertension (α_{34}) and COPD (α_{35}) are the weakest among the three clusters. Nonetheless, the results of our MoPTS model are in general agreement with the current understanding of the role of comorbidities in COVID-19 outcomes (CDCb, 2020).

While the geographical distribution of the clusters are no doubt represented by the populations therein and the underlying health and behavioral risk factors, environmental exposures, chronic conditions and comorbidities, it is important to avoid the risk of so-called ecological fallacy in attempting to infer individual disease outcomes based on data or results obtained at the level of large cities or counties. Indeed, the aim of our analysis is to provide insights into the existence of multiple patterns by which the current pandemic could affect the death rates in different metropolitan areas or counties in terms of their prevailing comorbidities. Such patterns could vary locally even within the same state. For instance, the disease dynamics as of 31 May 2020, at Allegheny county, which is located in western Pennsylvania and contains the city of Pittsburgh, was set apart (in cluster 3) from the other counties in the same state that lie more to the east and, indeed, share a different dynamic pattern (all of these are in cluster 2).

In the future, we aim to study possible interactions between different comorbidities as well as potential pathways leading from such comorbidities to various COVID-19 outcomes that were observed in diverse populations globally. With further availability of electronic health records, clinical evaluation and other microdata derived from monitoring the progression of the disease, more precise inferences can be drawn. Towards this, we could extend our present approach to time series models such as due to Hajrajabi and Maleki (2019) and Zarrin *et al*. (2019), and also the flexible class of distributions introduced by Hoseinzadeh *et al*. (2019), Moravveji *et al*. (2019) and Maleki *et al*. (2019).

As the pandemic progresses, we will have better understanding of the complex interplay of different comorbidities – acting either singly or in combination – in COVID-19 disease progression, especially of those leading to severe outcomes. Naturally, population differences between diverse geographies and societies might make such associations hard to generalize. Yet, certain putative risk factors, such as Type 2 diabetes or CHD, may be common among populations around the world. Toward this end, healthcare systems may want to develop pandemic resiliency with targeted communications and policies that are aimed at patients grouped by low to high risk categories based on their health, lifestyles, and environments. Such risk stratification would enable healthcare systems to prepare for effectively treating critical cases of the disease and minimizing fatality during both this pandemic as well as those in the future based on the known prevalence of comorbidities within the populations they serve.

References

Basford, K. E., Greenway, D. R., McLachlan, G. J. and Peel, D. (1997). Standard errors of fitted means under normal mixture models. *Computational Statistics*, **12**, 1-17.

CDCa. People who are at higher risk for severe illness. *Centers for Disease Control and Prevention.* https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-at-higher-risk.html.

CDCb. Information for healthcare professionals: COVID-19 and underlying conditions. *Centers for Disease Control and Prevention.* https://www.cdc.gov/coronavirus/2019-ncov/hcp/underlying-conditions.html.

CDCc. Behavioral risk factor surveillance system. *Centers for Disease Control and Prevention*. https://www.cdc.gov/brfss/data_documentation/index.htm.

Chen, J., Lu, H., Melino, G., et al. (2020). COVID-19 infection: the China and Italy perspectives. *Cell Death and Disease*. **11(6)**, 438. doi:10.1038/s41419-020-2603-0.

COVID-19 United States cases by county. *Johns Hopkins University and Medicine*. https://coronavirus.jhu.edu/us-map.

Cummings. M. J., Baldwin, M. R., Abrams, D., et al. (2020). Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: a prospective cohort study. *Lancet,* **395(10239)**, 1763-1770. doi:10.1016/S0140-6736(20)31189-2.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **B39**, 1–38.

Feng, Z., Li, Q., Zhang, Y., et al. (2020). The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. *China Centers for Disease Control and Prevention Weekly*, **1**, 145-151. doi:10.3760/cma.j.issn.0254-6450.2020.02.003.

Guan, W. J., Liang, W. H., Zhao, Y., et al. (2020). Comorbidity and its impact on 1590 patients with Covid-19 in China: A Nationwide Analysis. *European Respiratory Journal*, **55(5)**, 2000547. doi:10.1183/13993003.00547-2020.

Hajrajabi, A. and Maleki, M. (2019). Nonlinear semiparametric autoregressive model with finite mixtures of scale mixtures of skew normal innovations. *Journal of Applied Statistics,* **46(11)**, 2010-2029.

Hoseinzadeh, A., Maleki, M., Khodadadi, Z. and Contreras-Reyes, J. E. (2019). The skew-reflected-Gompertz distribution for analyzing symmetric and asymmetric data. *Journal of Computational and Applied Mathematics,* **349**, 132-141.

Jones, P. N. and McLachlan, G. J. (1992). Fitting finite mixture models in a regression context. *Australian Journal of Statistics,* **34**, 233–240.

Maleki, M., Mahmoudi, M. R., Wraith, D., Pho, K. H. (2019a). Time series modelling to forecast the confirmed and recovered cases of COVID-19. *Travel Medicine and Infectious Disease,* https://doi.org/10.1016/j.tmaid.2020.101742.

Maleki, M., Contreras-Reyes, J. E., Mahmoudi, M. R. (2019b). Robust mixture modeling based on two-piece scale mixtures of normal family. *Axioms*, **8(2)**, 38.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley. (ISBN: 978-0-471-00626-8).

Moravveji, M., Khodadadi, Z., Maleki, M. (2019). A Bayesian analysis of two-piece distributions based on the scale mixtures of normal family. *Iranian Journal of Science and Technology, Transactions A: Science*, **43(3)**, 991–1001.

Richardson, S., Hirsch, J. S., Narasimhan, M., et al. (2020). Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York city area. *Journal of the American Medical Association*, **10022**, 1-8. doi:10.1001/jama.2020.6775.

Statista. Italy: comorbidities in COVID-19 deceased patients 2020. https://www.statista.com/statistics/1110949/common-comorbidities-in-covid-19-deceased-patients-in-italy/.

U.S. Census Bureau. https://www.census.gov/quickfacts/.

Wang, B., Li, R., Lu, Z., Huang, Y. (2020). Does comorbidity increase the risk of patients with covid-19: Evidence from meta-analysis. *Aging*, **12(7)**, 6049-6057. doi:10.18632/AGING.103000.

Yang, J., Zheng, Y., Gou, X., Pu, K., Chen, Z., Guo, Q., et al. (2020). Prevalence of comorbidities and its effects in patients infected with SARS CoV-2: a systematic review and meta-analysis. *International Journal of Infectious Diseases,* **94**, 91-95. https://doi.org/10.1016/j.ijid.2020.03.017.

Zarrin, P., Maleki, M., Khodadadi, Z., Arellano-Valle, R. B. (2018). Time series process based on the unrestricted skew normal process. *Journal of Statistical Computation and Simulation*, **89** (**1**), 38-51.

APPENDIX

Table 1: Estimated parameters of the MoPTS $(\boldsymbol{D = 3, g = 3})$ with standard deviations.

$\boldsymbol{C_1}$	Est.	S.D.	$\boldsymbol{C_2}$	Est.	S.D.	$\boldsymbol{C_3}$	Est.	S.D.
$\boldsymbol{\pi_1}$	0.244186	4.235e-01	π_2	0.33721	3.683e-02	π_3	0.418604	5.012e-02
α_{11}	2.09397	8.311e-02	α_{21}	0.54520	1.321e-01	α_{31}	–0.02475	1.074e-03
$\boldsymbol{\alpha_{12}}$	–1.32063	3.242e-02	α_{22}	–0.13055	2.032e-02	α_{32}	–0.03225	2.034e-02
$\boldsymbol{\alpha_{13}}$	–2.66249	4.101e-01	α_{23}	0.65799	1.721e-02	α_{33}	–0.02447	2.984e-03
$\boldsymbol{\alpha_{14}}$	1.62983	3.422e-03	α_{24}	–0.14953	2.032e-03	α_{34}	0.06790	3.352e-03
$\boldsymbol{\alpha_{15}}$	–0.29323	3.012e-02	α_{25}	0.52714	2.857e-02	α_{35}	0.03763	2.405e-02
$\boldsymbol{\beta_{1,1}}$	–1.85318	6.037e-02	$\beta_{2,1}$	–0.73512	9.063e-02	$\beta_{3,1}$	–0.08560	7.342e-03
$\boldsymbol{\beta_{1,2}}$	0.14396	3.755e-02	$\beta_{2,2}$	0.048204	4.234e-03	$\beta_{3,2}$	0.00883	8.311e-03
$\boldsymbol{\beta_{1,3}}$	–0.00136	4.937e-05	$\beta_{2,3}$	–0.00039	9.043e-04	$\beta_{3,3}$	–0.00006	1.003e-05
$\boldsymbol{\sigma_1^2}$	574.38301	2.223e01	σ_2^2	75.21687	0.389e01	σ_3^2	16.30303	0.232e01

Table 2: Three clusters of 86 U.S. cities identified by the 3 components of the MoPTS model. The county and the state (including District of Columbia) of each city are shown.

Cluster 1 (21 members)	**Cluster 2** (29 members)	**Cluster 3** (36 members)
Fairfield, Connecticut	Arapahoe, Colorado	Jefferson, Alabama
Hartford, Connecticut	Denver, Colorado	Mobile, Alabama
New Haven, Connecticut	Weld, Colorado	Maricopa, Arizona
Dougherty, Georgia	Washington, District of Columbia	Pima, Arizona
Jefferson, Louisiana	Cook, Illinois	Los Angeles, California
Orleans, Louisiana	DuPage, Illinois	Orange, California
Essex, Massachusetts	Lake, Illinois	Riverside, California
Hampden, Massachusetts	Will, Illinois	San Bernardino, California
Middlesex, Massachusetts	Lake, Indiana	San Diego, California
Norfolk, Massachusetts	Marion, Indiana	Santa Clara, California
Plymouth, Massachusetts	Caddo, Louisiana	Adams, Colorado
Suffolk, Massachusetts	East Baton Rouge, Louisiana	Jefferson, Colorado
Macomb, Michigan	Baltimore City, Maryland	New Castle, Delaware
Oakland, Michigan	Bristol, Massachusetts	Broward, Florida
Wayne, Michigan	Worcester, Massachusetts	Lee, Florida
Essex, New Jersey	Genesee, Michigan	Miami Dade, Florida
Hudson, New Jersey	Hennepin, Minnesota	Palm Beach, Florida
Mercer, New Jersey	St. Louis City, Missouri	DeKalb, Georgia
Passaic, New Jersey	Camden, New Jersey	Fulton, Georgia
Union, New Jersey	Erie, New York	Kane, Illinois
Westchester, New York	Lucas, Ohio	Polk, Iowa
	Mahoning, Ohio	Jefferson, Kentucky
	Berks, Pennsylvania	Ramsey, Minnesota
	Lackawanna, Pennsylvania	Clark, Nevada
	Lehigh, Pennsylvania	Hillsborough, New Hampshire
	Northampton, Pennsylvania	Onondaga, New York
	Philadelphia, Pennsylvania	Cuyahoga, Ohio
	King, Washington	Franklin, Ohio
	Milwaukee, Wisconsin	Hamilton, Ohio
		Summit, Ohio
		Allegheny, Pennsylvania
		Shelby, Tennessee
		Dallas, Texas
		Harris, Texas
		Tarrant, Texas
		Snohomish, Washington

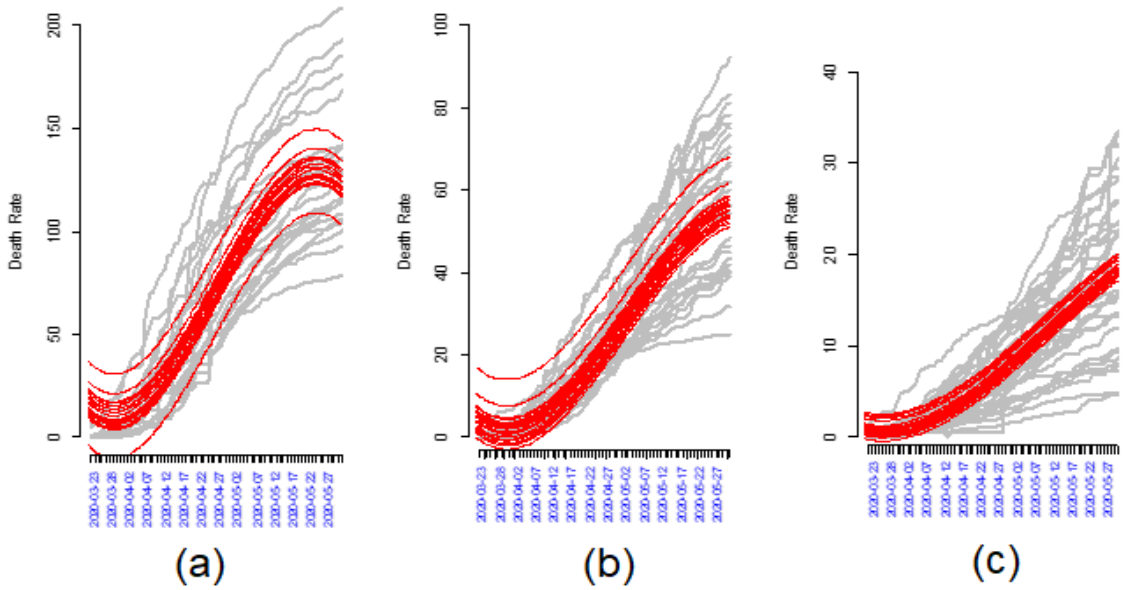

Figure 1: Time series plots (grey) of 86 U.S. cities belonging to clusters 1 (a), 2 (b) and 3 (c) based on $g = 3$ mixture components, superimposed with the fitted MoPTS model (red) for each city. The x-axis denotes time and y-axis COVID-19 death rate. The clusters show marked differences in their respective ranges of death rates (y-axes) over time.

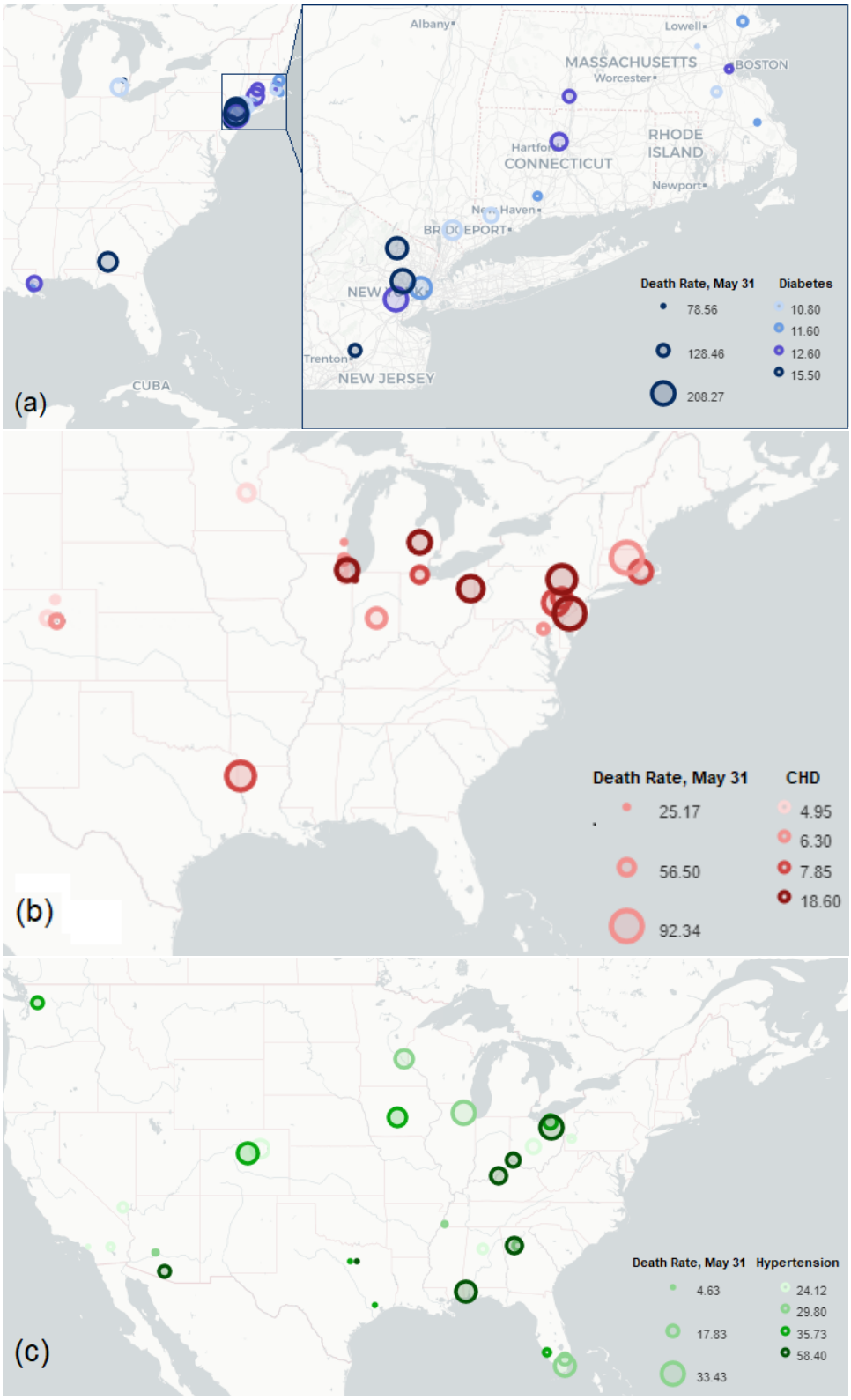

Figure 2. Geographical distribution of the U.S. cities in clusters 1 (a), 2 (b), and 3 (c) based on MoPTS model. On a map, a bubble's location shows the latitude and longitude of a city *C*, its size is proportional to the COVID-19-specific death rate at *C* as on May 31, 2020, while its shading is proportional to the prevalence of the leading comorbidity in the cluster to which *C* belongs.