Point Process Models for COVID-19 Cases and Deaths

Álvaro Gajardo [1] and Hans-Georg Müller[1]

[1] Department of Statistics, University of California, Davis, CA 95616 U.S.A.

ARTICLE HISTORY
Compiled February 19, 2023

ABSTRACT
The study of events distributed over time which can be quantified as point processes has attracted much interest over the years due to its wide range of applications. It has recently gained new relevance due to the COVID-19 case and death processes associated with SARS-CoV-2 case processes that characterize the COVID-19 pandemic and are observed across different countries. It is of interest to study the behavior of these point processes and how they may be related to covariates such as mobility restrictions, gross domestic product per capita, and fraction of population of older age. As infections and deaths in a region are intrinsically events that arrive at random times, a point process approach is natural for this setting. We adopt techniques for conditional functional point processes that target point processes as responses with vector covariates as predictors, and extend it to study the interaction and optimal transport between case and death processes and doubling times conditional on covariates.

KEYWORDS
Fréchet Regression, Intensity Function, Optimal Transport, Wasserstein Metric, SARS-CoV-2, Poisson Process, Cox Process.

1. Introduction

The outbreak of the COVID-19 pandemic has impacted most countries around the world due in part to the strong connectivity of the global network as well as intrinsic characteristics of the virus, especially its infectivity, easily spreading from person to person by day-to-day interactions, especially indoors and under conditions of prolonged contact, and the fact that asymptomatic subjects can be highly contagious. The total number of worldwide confirmed cases has reached around 5.4 million people with more than 345 thousand deaths by May 24, 2020, with continuously rising numbers of infections and deaths [19]. It is of great urgency to understand how the time evolution of the case and death counts is associated with factors such as mitigation or social distancing measures, wealth and age of the population of a country, among other factors. We address this problem by a conditional point process model, which provides a natural framework in this setting, since confirmed infections (cases) and deaths due to COVID-19 are events which arrive at random times within each region of interest, i.e., they come from a temporal point process mechanism [10] for each region (country).

Approaches that take into account the integer-valued character of the number of confirmed cases and deaths have been explored in [20], where standard Poisson and Negative Binomial Generalized Linear Models have been applied to study the effect

of covariates on the daily case counts in China. This approach takes the form of the data as daily counts into account, but does not allow to study the effect of covariates on the distribution of the infection events over time, as it only models the total case counts per day, and it also does not incorporate information from different countries, which is essential to arrive at more general conclusions. In related work, [23] applied Poisson regression and Generalized Additive Models (GAM) to the study of the basic reproduction number R_0, which can be interpreted as the expected number of infections directly generated from one case when the entire population is at risk, while [28] employed GAM and a quasi-Poisson approach to estimate the doubling number across countries and [31] utilized time series analysis to forecast the worldwide number of cases. Poisson mixture models were applied by [22] to study the time varying case fatality rate in China, while [34] fitted a generalized logistic growth model to confirmed case curves. These latter approaches do not incorporate covariates, are based on continuous approximations to case curves and ignore the point process aspects.

Our goal in this paper is to develop a point process perspective with the additional goal to study the interaction between the case count and death count point processes. We consider the confirmed COVID-19 infections or deaths that occur in a given country during a time window $[0, T]$, $T > 0$, for some suitable chosen initial time $t = 0$ that may be region-specific, and which we choose as the first time when a country records 80 cases of confirmed infections. We then view infections (or deaths) as events that occur at random times $T_1, \ldots, T_{N_i}$ in the interval $[0, T]$ for the i-th country, which are the data associated with the point process approach. Figure 1 shows the infection and death point process for $n = 62$ countries that we consider in this study. The selection of an initial time when 80+ cases are recorded provides for a temporal alignment of the processes as the pandemic reaches the countries at different calendar times so that the calendar time scale is not meaningful for the point process modeling perspective.

A key quantity of interest is the local intensity function which can be interpreted as the expected number of events per unit time. In the context of doubly stochastic Poisson or Cox processes [9] that we consider here, the intensity function $\Lambda(t)$ is assumed to be random and such that conditional on a realization $\Lambda = \lambda$, $N(t)$ is a non-homogeneous Poisson process with intensity $\lambda(t)$. The function $\Lambda(t)$ characterizes the point process and viewing it as a random function provides the flexibility to include various countries in the study, as each country's underlying intensity function that characterizes the country-specific point process can be thought of as a realization of an underlying stochastic intensity process.

The available COVID-19 data is usually recorded at the daily level, where the total counts per day are recorded instead of each individual event time [19]. This poses further challenges that we address by regarding the observed data as a binned version of the actual event times, corresponding to pre-smoothing step that is the result of binning over each day. Such binned versions of a point process can be viewed as exhibiting daily granularity [17]. Functional Data Analysis (FDA) [18, 33] suggests an approach based on approximating the binned daily counts for each country (or monotone transformations of the counts) by smooth trajectories and then applying FDA techniques to analyze the time dynamics of accumulated case or death curves [5]. This incorporates and pools the available information across different countries and allows to study the main modes of variation for the curves [7], but does not reflect the underlying monotonicity of the cumulative case and death curves, which leads to a nonlinearity of the subspace of squared integrable functions where these curves are assumed to live [29].

The dynamic modeling of key epidemiological quantities such as the basic repro-

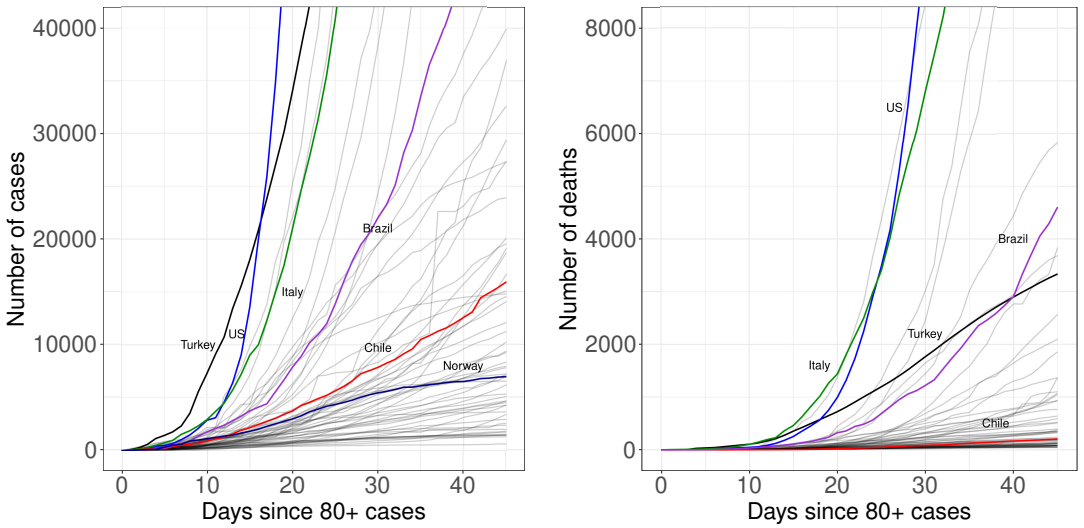

Figure 1. Infection and death point processes viewed as trajectories of cumulative cases and deaths for several countries over a time window of 45 days since reaching 80+ cases. This alignment serves to make the temporal evolution of the point processes comparable as the pandemic reached each country at a different calendar time.

duction number R_0 or doubling time is also of great interest, where [24] adopted a Hawkes point process approach to analyze the worldwide and country-wise time varying reproduction number $R(t)$, the reproduction number at time t. For predicting $R(t)$ in Wuhan, China, [21] employed a dynamic transmission stochastic model, while [35] proposed time varying parametric models to explain the evolution of the confirmed cases, followed by estimating the doubling number and $R(t)$. In a similar approach, [1] employed SIDR models to estimate $R(t)$ and to then forecast the spread of the virus over time. The relation of covariates such as age, gender or mobility on the mortality and the spread of the virus has also been explored [4, 11].

We adopt here a functional non-parametric conditional point process approach [13] to study the association between vector covariates and the infection and death point processes as responses, and extend it to study the relation between the arrival times of these two paired processes in the presence of covariates along with key quantities of interest such as the doubling time. Details about data sources can be found in Section 2.1 below.

2. Methods

2.1. Data

We obtained the infection and death information for each country from the COVID-19 Data Repository at the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, which is publicly available at `https://github.com/CSSEGISandData/COVID-19`. This database is updated on a daily basis and contains the cumulative number of confirmed cases and deaths for each country from Jan 22, 2020, and was accessed on Jan 6, 2021. We consider the point process of infections or deaths per country over a time window of 45 days since the first up-crossing of 80 confirmed cases. This allows to compare countries over comparable time windows, as the pandemic reached countries at different times and this impedes comparisons based

on calendar time.

We further included only countries that reported at least 50 days of non-zero death counts after the first time that 80 confirmed cases are reached and whose total death counts are at least 40 at the end of this time window. This allows to remove right boundary effects as explained in Section 2.2. Since for some countries the cumulative case and death curves were not monotonically increasing due to potential data accuracy issues at some dates, we enforced monotonicity by keeping the cumulative counts at the same level as the previous day for those dates where it decreased. We did not consider the countries South Korea and Thailand due to their outlying pattern in their daily new case counts as they presented a strong decay in new daily cases from early on. The total number of countries in the study is $n = 62$.

We obtained the Gross Domestic Product (GDP) per capita in 2018 for each included country from the worldbank database, which is publicly available at `https://data.worldbank.org`. The Google mobility report [16] contains the daily percent change of visits to different place categories such as workplace, residential, grocery, among others, as compared to a baseline which is defined as the median value of the corresponding day of the week during the period Jan 3 to Feb 6, 2020. The data is publicly available at `https://www.google.com/covid19/mobility/` and was accessed on Jan 6, 2021. We work with the Google mobility report for workplaces since this is a natural proxy for infection chances due to person to person interactions. Denoting by $\Delta_i(t)$ the workplace mobility on day t for country i, $i = 1, \ldots, n$, we construct the integrated workplace mobility covariate as $\int_0^T \Delta_i(t)dt$, which is numerically approximated and is referred to as the mobility index. Thus, higher values of the mobility index reflect higher chances for infection. We use the R package fdapace to perform the numerical integration [6]. The ranges of the covariates and the positioning of the countries can be seen in Figure 6.

2.2. Local Fréchet regression for point processes as responses

Let $N(t)$ denote a generic temporal point process that represents the number of events that occur in a time window $[0, t]$, where $t \leq T$ for some endpoint $T > 0$. The local intensity function of the process $N(t)$ is defined as

$$\lim_{\Delta t \to 0} \frac{E(N(t + \Delta t)) - E(N(t))}{\Delta t}, \tag{1}$$

and can be interpreted as the rate of occurrence of events per unit time. We consider $N(t)$ to be a Cox process [9]. Cox processes are characterized by the existence of an underlying stochastic positive and integrable intensity function Λ such that, conditional on a realization $\Lambda = \lambda$, $N(t)$ follows a non-homogeneous Poisson process with intensity function $\lambda(t)$. In practice, the function $\Lambda(\cdot)$ remains unobserved as only the arrival times $T_1, \ldots, T_{N(T)}$ of the point process $N(t)$ are available. A key property that connects the arrival times to the underlying stochastic intensity function Λ is that, conditional on observing $N(T) = m > 0$ events and a realization $\Lambda = \lambda$, it holds that $T_1, \ldots, T_m \overset{iid}{\sim} f$, where $f(t) = \lambda(t)/\tau$ is a density function and $\tau = \int_0^T \lambda(s)ds$ is a scalar [10], also referred to as intensity factor. Point process techniques have recently also been successfully applied for bike rentals as events, aiming at the analysis of repeated observations of the bike rental point process [15], and the spatial distribution of street robberies [14].

The target is to study a notion of regression for the infinite dimensional object Λ, which characterizes the Cox process and lies in a suitable metric space, with the Euclidean covariates $X \in \mathbb{R}^p$, $p > 0$ as predictors. For this, we employ the recently proposed framework of Fréchet regression for objects that reside in general metric spaces, which is based on the concept of conditional Fréchet means. This is a generalization of the standard regression function $E(Y|X = x)$ for real-valued responses Y [30] and was extended recently to the point process setting in [13]. Denoting by (Ω, d) the metric space of intensity functions, the Fréchet regression of the random intensity $Y \in \Omega$ on the covariate $X = x$ is defined as

$$Y_{\oplus}(x) := \underset{\lambda \in \Omega}{\arg\min}\ E(d^2(Y, \lambda)|X = x). \tag{2}$$

Since intensity functions $Y \in \Omega$ can be written uniquely as a product between the density $f(t) = Y(t)/\int_0^T Y(s)ds$ and the scalar $\tau = \int_0^T Y(s)ds$, the intensity space Ω can be viewed as a product metric space $\Omega = \mathcal{D} \times \Omega_s$, where $\mathcal{D}$ and $\Omega_s = (0, \infty)$ denote the spaces of density functions over $[0, T]$ and intensity factors, respectively. We utilize the l^2 type product metric d between intensities $\Lambda_1 = (f_1, \tau_1)$ and $\Lambda_2 = (f_2, \tau_2)$, which is given by

$$d(\Lambda_1, \Lambda_2) = \left(d_W^2(f_1, f_2) + d_E^2(\tau_1, \tau_2)\right)^{1/2}, \tag{3}$$

where d_E is the Euclidean metric and d_W is the Wasserstein metric between probability distributions, which corresponds to the L^2 distance between the quantile functions associated with the densities f_1 and f_2. Since we measure differences in the intensity factor from changes in the density separately, the scale of the metrics is not relevant, i.e. the Fréchet regression function remains the same for all weighted metrics $d^2 = \alpha d_W^2 + \beta d_E^2$, as long as $\alpha, \beta > 0$ [13].

This allows to quantify differences in shapes and in the intensity factors separately since it can be shown that the intensity regression function is given by

$$\begin{aligned} Y_{\oplus}(t, x) &= f_{\oplus}(t, x)\tau_{\oplus}(x), \quad t \in [0, T], \\ \tau_{\oplus}(x) &= \max\{E(\tau|X = x), 0\}, \\ f_{\oplus}(\cdot, x) &= \underset{g \in \mathcal{D}}{\arg\inf}\ E(d_{\mathcal{W}}^2(f, g)|X = x). \end{aligned} \tag{4}$$

We refer to $f_{\oplus}(\cdot, x)$ as the density regression function, which measures how the shape of the arrival times is affected by the covariate level. For the estimation of $Y_{\oplus}(t, x)$, we consider replications of the point process as described below.

Let $X_i \in \mathbb{R}^p$ be the covariate vector of country i, $i = 1, \ldots, n$, where we have $p = 2$, since we consider GDP per capita and the mobility index as covariates. Denote by $N_i(t)$ the point process of either cases or deaths for the i-th country. We consider a replicated point process framework where $(X_i, \Lambda_i, N_i) \overset{iid}{\sim} (X, \Lambda, N)$, $i = 1, \ldots, n$, are the replications of the underlying COVID-19 count process (cases or deaths). The estimation is based on local multivariate polynomial regression methods [12], for which we introduce the well known local weights $s_{in}(x, h) = \frac{1}{\hat{\sigma}_0^2}[1 - \hat{u}_1^T \hat{u}_2^{-1}(X_i - x)]K_h(X_i - x)$, where $\hat{u}_j = n^{-1}\sum_{i=1}^n K_h(X_i - x)e_j(X_i)$, with $j \in \{0, 1, 2\}$, $e_0(X_i) = 1$, $e_1(X_i) = X_i - x$ and $e_2(X_i) = (X_i - x)(X_i - x)^T$, $\hat{\sigma}_0^2 = \hat{u}_0 - \hat{u}_1^T \hat{u}_2^{-1}\hat{u}_1$, $K_h(X_i - x) := \prod_{j=1}^p h_j^{-1} K((X_{ij} - x)/h)$, with X_{ij} being the jth coordinate of X_i. Here the kernel K is a continuous and

symmetric density function, and the h_j are a sequence of positive bandwidths. Then the estimator of $\tau_\oplus(x)$ is given by

$$\hat{\tau}_\oplus(x) = \max\left(0, \frac{n^{-1}\sum_{i=1}^{n} s_{in}(x,h)N_i(T)}{\bar{N}(T)}\right), \tag{5}$$

where the ratio by the average country-wise total counts is to ensure stability of the estimator in an asymptotic infill point process framework [26] that allows consistent estimation of the population quantities [13].

For the estimation of the density regression part $f_\oplus(\cdot, x)$, we start with an estimate $\hat{Q}_i$ of the quantile function associated with the arrival times of the process N_i. Then an estimate $\tilde{f}_\oplus(\cdot, x)$ of $f_\oplus(\cdot, x)$ is constructed as follows: First, the empirical estimate of the quantile function $\tilde{Q}_\oplus(\cdot, x)$ corresponding to the density $\tilde{f}_\oplus(\cdot, x)$ is obtained by minimizing

$$\tilde{Q}_\oplus(\cdot, x) := \underset{q\in\mathcal{Q}}{\arg\min}\ ||q - n^{-1}\sum_{i=1}^{n} s_{in}(x,h)\hat{Q}_i||^2_{L^2([0,1])}, \tag{6}$$

where $\mathcal{Q}$ is the space of quantile functions which are (M, L) bi-Lipschitz, $M, L > 0$. This ensures that the underlying densities are smooth and bounded away from zero. We numerically solve (6) by casting it as a quadratic optimization problem along with the constraints, and utilizing very small and large constants $M = 10^{-10}$ and $L = M^{-1}$ [13]. Finally, $\tilde{f}_\oplus(\cdot, x)$ is obtained by mapping the quantile $\tilde{Q}_\oplus(\cdot, x)$ to density space. We chose K as a Gaussian kernel and the bandwidths as 20% of each covariate range which worked well to capture the underlying patterns. For several of these steps we use the Fréchet R package [8], which is available on Github at `https://github.com/functionaldata/tFrechet`.

In practice, for both the infection and death point processes, the available data does not contain the exact time of occurrence of each event as national health institutions and governments report the corresponding events aggregated on a daily basis. Thus, the arrival times are not directly available in the continuum but rather as binned data over a fine daily grid. We regard this as a form of pre-smoothing and utilize a kernel density smoother for binned data with equally spaced bins [17] to first obtain the estimated density function $\hat{f}_i$ associated with each country i, $i = 1, \ldots, n$. These density function estimates are then mapped to quantile space to obtain $\hat{Q}_i$. Further details can be found in the Supplement.

Finally, the estimated intensity regression function is given by

$$\hat{Y}_\oplus(t, x) = \tilde{f}_\oplus(\cdot, x)\hat{\tau}_\oplus(x), \tag{7}$$

which was shown to converge to its true counterpart up to the constant $E(\tau)$ in an asymptotic infill framework [13], so that the relative ratios, which are of central importance, are on target.

Since local polynomial regression techniques are utilized for the two-dimensional predictor X, we only show the estimated regression functions over a region strictly contained in the range of the covariates in order to avoid boundary effects, which otherwise often dominate the estimation error [25]. Similarly, to avoid boundary effects of the density estimator $\hat{f}_i$ near the right endpoint, we considered an extended time window of 50 days in the estimation step and then display the estimate over $[0, 45]$

by re-normalizing the density estimates. To mitigate boundary effects near the left endpoint, we considered the data before the first up-crossing of 80 cases which provides information to the left of the domain starting point, thus reducing boundary effects.

2.3. Time and covariate dependent doubling time

Consider $N_{\oplus}(t)$ to be a non-homogeneous Poisson process with intensity function $\lambda_{\oplus}(\cdot, x)$, which corresponds to the intensity regression function at covariate level $X = x \in \mathbb{R}^2$ for either the infection or death process. We define $\vartheta = \vartheta(t, x)$ as the doubling time of $N_{\oplus}(t)$ at time t and predictor level x if

$$E[N_{\oplus}(t + \vartheta)] = 2E[N_{\oplus}(t)]. \tag{8}$$

It is then of interest to study how the doubling time ϑ varies with t and the covariates X. From the Poisson assumption on $N_{\oplus}(t)$, it follows that (8) is equivalent to

$$\int_0^{t+\vartheta} \lambda_{\oplus}(s, x)ds = 2\int_0^t \lambda_{\oplus}(s, x)ds. \tag{9}$$

Thus, $\vartheta(t, x)$ can be found by solving (9) numerically over a smaller time window provided that there exists a solution.

We chose the times $t \in [1, 24]$ days to carry out the numerical scheme, where we start after $t = 0$ due to sparseness of the death process in that region. Note that even though the estimation of the intensity regression function can be achieved up to the constant $E(\tau)$, this factor drops out in the previous relation, so that estimation of ϑ can proceed as if one had a consistent estimate of the intensity factor $E(\tau)$.

2.4. Optimal transport between case and death processes

Let $N^c_{\oplus}(t)$ and $N^d_{\oplus}(t)$ be the case and death regression point processes with intensity function $\lambda^c_{\oplus}(t, x) = f^c_{\oplus}(t, x)\tau^c_{\oplus}(x)$ and $\lambda^d_{\oplus}(t, x) = f^d_{\oplus}(t, x)\tau^d_{\oplus}(x)$, respectively, as described in Section 2.3. Suppose that T_c and T_d are random variables with density functions $f^c_{\oplus}(\cdot, x)$ and $f^d_{\oplus}(\cdot, x)$, which represent random arrival times from the case or death process conditional on $X = x$, respectively. The optimal transport plan in the sense of Monge [32] from the case to the death process for a given covariate level $X = x$ is defined as the map $\psi(t) = T^{\oplus}_{c\to d}(t, x) : [0, T] \to [0, T]$ that pushes the distribution of the case arrival times to that of the death process by minimizing the transport cost

$$T^{\oplus}_{c\to d}(t, x) = \underset{T^{\oplus}:[0,T]\to[0,T]}{\operatorname{arginf}} \int_0^T (s - T^{\oplus}(s))^2 f^c_{\oplus}(s, x)ds, \tag{10}$$

which is also known as the optimal transport between the probability measures corresponding to the densities $f^c_{\oplus}(\cdot, x)$ and $f^d_{\oplus}(\cdot, x)$ under the 2-Wasserstein metric. The solution to (10) is well known and is given by $T_{c\to d}(t, x) = Q^d_{\oplus}(F^c_{\oplus}(t, x), x)$, where $Q^d_{\oplus}(\cdot, x)$ is the quantile function corresponding to the density $f^d_{\oplus}(\cdot, x)$ and $F^c_{\oplus}(\cdot, x)$ is the distribution function associated with the density $f^c_{\oplus}(\cdot, x)$.

To quantify the transport specifically for the transition from cases to deaths, we

consider the distance measure

$$\rho_{\oplus}(x) = \int_0^T (T^{\oplus}_{c\to d}(t,x) - t) 1_{\{T^{\oplus}_{c\to d}(t,x)\geq t\}} dt,$$

where $x \in \mathbb{R}^2$ is the covariate level. This quantity can be interpreted as a measure of how farther away the deaths occur as compared to the cases when $X = x$, i.e., it serves as a proxy of the underlying time lag.

We compute $\rho_{\oplus}(X_i)$ where X_i are the covariate levels of country i, $i = 1, \ldots, n$, and test the joint significance of the covariates. The latter is done under a classical multivariate linear regression setting with $\rho_{\oplus}(X)$ as response, where we include the main and squared covariate effects as predictors, i.e., $E(\rho_{\oplus}|\mathbf{x_1}, \mathbf{x_2}) = \beta_0 + \beta_1 \mathbf{x_1} + \beta_2 \mathbf{x_2} + \beta_3 \mathbf{x_1^2} + \beta_4 \mathbf{x_2^2}$, where $\mathbf{x_1}$ and $\mathbf{x_2}$ denote the mobility index and GDP per capita covariates, respectively. Thus, the effect of one covariate on $\rho_{\oplus}$ depends on the level of the other as is showcased in Figure **??** in the Supplement. For more details we refer to the Supplement.

3. Results

3.1. COVID-19 case and death point processes

3.1.1. Conditional COVID-19 point processes

We regress the case point process on the integrated Google mobility data and the Gross Domestic Product per capita (see Data 2.1). It is of interest to study the relation of the case intensity function with each covariate, which we address by varying one covariate while keeping the other covariate at its mean level. Figure 2 shows that decreased levels of the mobility index are associated with a uniformly lower intensity function over the entire time window $[0, 45]$ days since the first time the pandemic hits 80 or more cases, which is likely due to less opportunity for the virus to spread through person to person interactions as mobility decreases. A similar pattern of accelerated cases and deaths is observed when GDP per capita increases while keeping the mobility index at its mean level. This may indicate that there were more opportunities for infection in richer countries during the time domain that is considered but it could also be a consequence of differences in testing for the virus.

The intensity function increases towards a unique peak, where the location of the peak moves to the right with increasing mobility index while it moves to the left with increasing GDP per capita. Specifically, the density function of the case arrival times moves to the left with increasing GDP per capita while the mobility index is kept at the mean level, suggesting that overall, cases are more likely to occur sooner rather than later for richer countries under a similar level of the mobility index. Moreover, higher levels of the mobility index are seen to be associated with cases located increasingly towards the end of the time window rather than at its beginning. This may be related to the incubation period that follows after a person has been infected until symptoms appear or testing is performed, all of which may lead to a time lag. Similar results are observed for the intensity functions of the death process conditional on the covariates. (see Figure **??** in the Supplement).

Here the integrated Google mobility index $\int_0^T \Delta(t)dt$ corresponds to an overall measure of mobility. Alternatively, one could consider regressing the cases or death point

process on the entire (functional) infinite-dimensional predictor $\Delta(t)$. We found that the latter approach is equivalent to regressing the point process on the first functional principal component of $\Delta(t)$ as given by its Karhunen-Loève representation; details on this are provided in the Supplement.

3.1.2. Doubling time dynamics

The propagation of the virus in time can be measured by how long it takes, on average, to double the current number of cases at time t. We refer to this quantity as the doubling time ϑ. (see Methods 2.3). Lower values indicate faster spread of the virus while its rate of change measures the instantaneous severity of the disease in terms of propagation. Figure 3 shows that decreased opportunities for infection and the corresponding lower levels of mobility in the population are associated with increased doubling times over an initial time window of 24 days since the initial spread of 80 or more cases.

Moreover, for a country with a mean GDP per capita level and mobility index around the 75% quantile, the doubling time appears to increase exponentially which suggests the long term positive impact of containment policies. In fact, by the 15^{th} day since the initial spread, roughly eight more days are needed on average for the cases to duplicate while the doubling requires more than 17 days nine days later.

Higher GDP per capita levels appear to be associated with shorter doubling times on average and thus higher risk of propagation of the pandemic. We observe an inflection point around $t = 18$ days, where the doubling time is higher for wealthier countries, which may be explained by a lagged effect of closing borders and the change from physical to remote business meetings. Similar results hold for the doubling time of the death process (see Figure **??** in the Supplement).

3.2. From case to death process: An optimal transport approach

Although studying the individual relation between the case or death processes and the covariates is important, exploring the interactions between both processes in the presence of predictors is also of interest as they are naturally connected in that infections precede deaths and the patterns of the time lag between these contributes to our understanding of the progression of the pandemic. To study these patterns of time delay, we employ optimal transport techniques, adapted to the bivariate point process setting (see Methods 2.4).

We consider the case point process at covariate level $x \in \mathbb{R}^2$, which is a Poisson point process with intensity function $\lambda^c_\oplus(\cdot, x)$. Similarly, for the death process we consider the intensity to be $\lambda^d_\oplus(\cdot, x)$. The interaction between these paired point processes can be measured by the optimal transport plan from the case to the death arrivals times, which for probability distributions is defined as moving mass from one distribution to the other under the constraint that the total transport cost is minimized (see Methods 2.4), and which is closely connected with the Wasserstein metric between probability distributions. The optimal transport and Wasserstein perspective has been very successful in statistical modeling [2, 3, 27, 36].

Increased levels of GDP per capita, while keeping the mobility index at the mean, translates into a uniformly higher optimal transport map, where transports move further to the right, from the case to the death process, as can be seen in Figure 4. Thus, wealthier countries appear to have a much larger lag between deaths and infections. Likely this is because wealthier countries have better infrastructure and

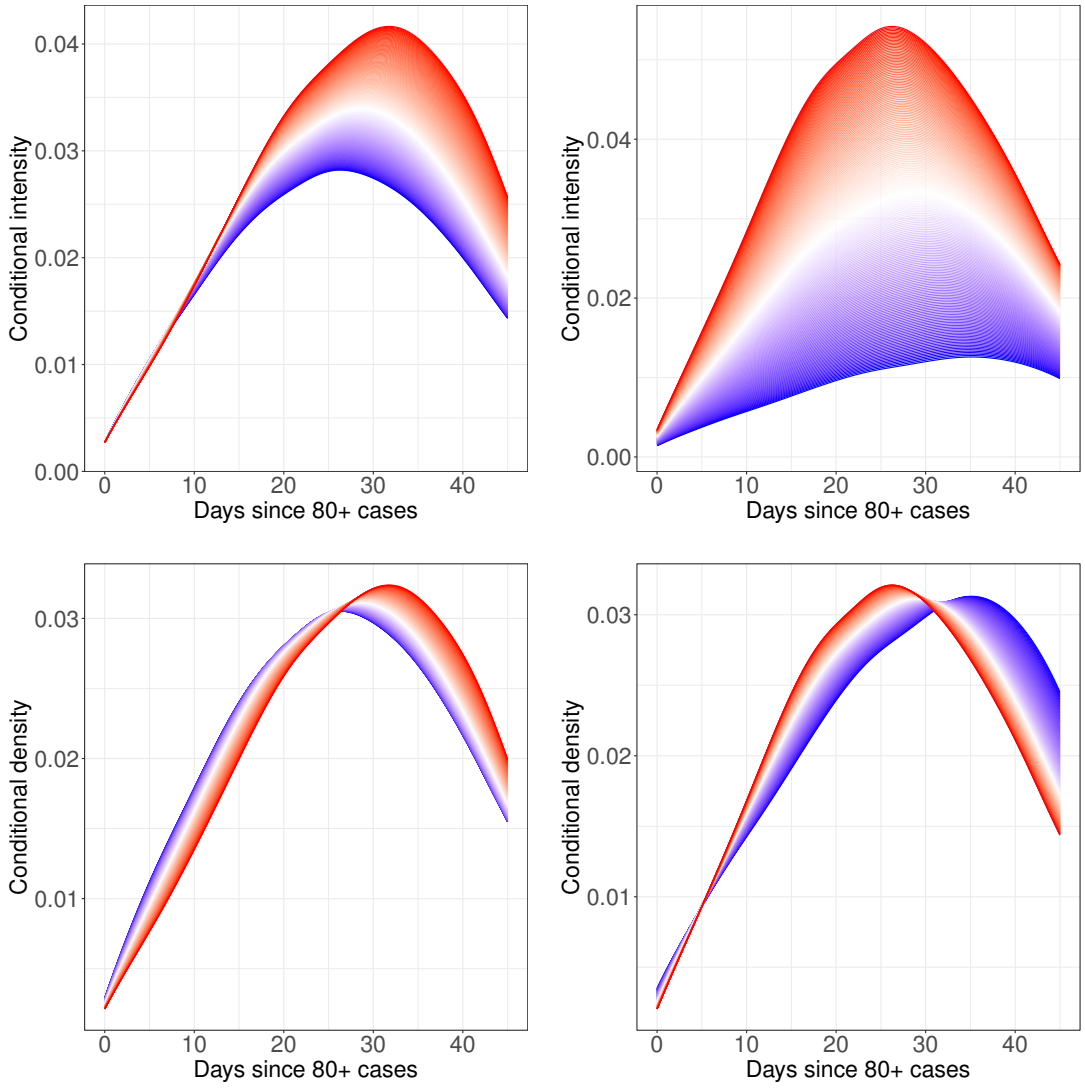

Figure 2. Local Fréchet regression estimates for the intensity (top) and density (bottom) functions for the point process of cases. The left panels show the situation of increasing the mobility index from the 25% to 75% quantiles (blue to red), i.e., increased mobility, while keeping the GDP per capita at the mean level. The right panels display the effect of increasing GDP per capita from the 25% to 75% quantile (blue to red) while keeping mobility at the mean level.

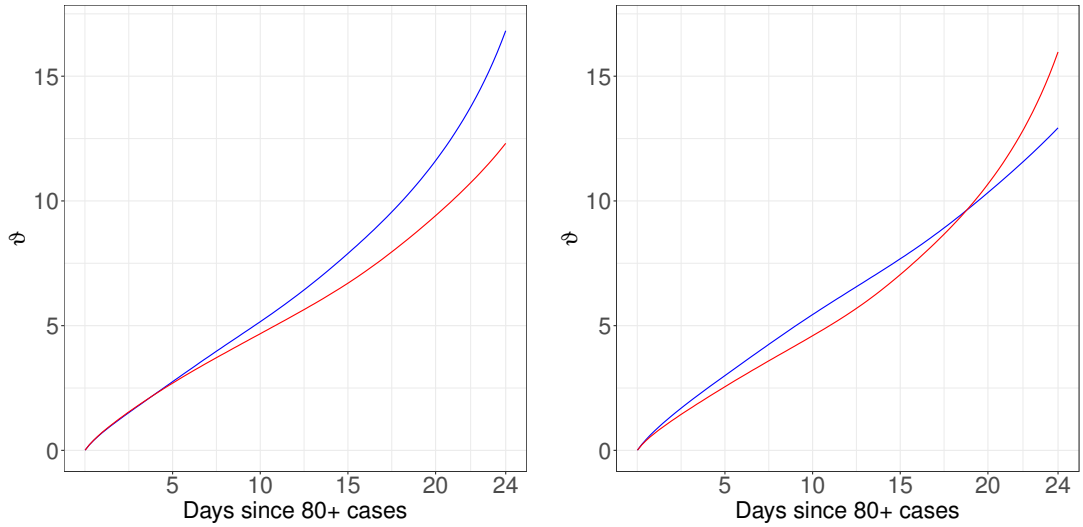

Figure 3. Doubling times for the case point process under different covariate levels. The left panel shows the estimate when the mobility index is at the 25% and 75% quantiles (blue and red, resp.) while keeping the GDP per capita at mean level. The right panel displays the estimate when increasing GDP per capita from the 25% to 75% quantile (blue to red) while keeping mobility at the mean level.

resources for the health system compared to countries with lower GDP per capita, and that the availability of health care resources determines whether death of critically ill patients can be postponed.

It is instructive to compare the density functions corresponding to the arrival times of the case and death point processes for Chile and Norway. Both countries have a mobility index around the mean level, but are very far apart in terms of wealth: Norway's GDP per capita is around five times that of Chile. The optimal transport result predicts that the death arrival times should be much more pushed to the right relative to the case arrival times for Norway as compared to Chile. That this is indeed the case is seen in Figure 5, where for Norway the cases and death arrival times are both unimodal with peaks around 19 and 30 days, respectively, while for Chile the deaths are pushed to the right with a higher density in the time interval $t \in [14, 38]$ days as compared to the cases density.

Moreover, increases in the mobility index (more opportunities for infection) are related to an overall lower optimal transport from cases to deaths which could be indicative of an overloaded health system, as higher mobility translates to a larger fraction of the population being infected and requiring medical assistance, under a situation of stretched resources. To assess the covariate effect on the interaction between the processes, we computed the area $\rho_{\oplus}(x)$ between the transport plan and the identity map, which is a measure of how farther away in time deaths occur as compared to cases and thus serves as a proxy of the natural temporal lag between these; the value is indicated for a sample of countries in Figure 5. We test for joint significance of the two covariates in a classical multivariate linear regression setting with main effects as well as quadratic terms for $\rho_{\oplus}(x)$ as response (see Methods 2.4) and obtain that the null hypothesis of no-covariate effect is rejected at $p < 2.2e{-}16$.

Figure 6 shows the positioning of all considered countries in terms of the covariates GDP per capita and mobility index, where each country is colored according to its $\rho_{\oplus}(x)$ value. As explained before, wealthier countries have higher values of $\rho_{\oplus}(x)$ which indicate increased time lags of deaths.

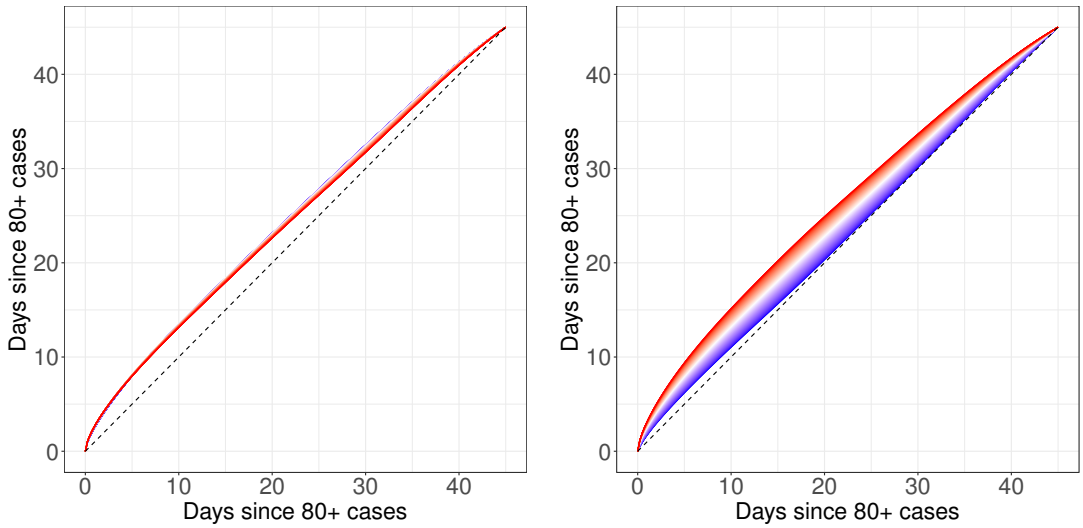

Figure 4. Optimal transport maps from case to death processes. The left panel shows the transport when the mobility index is increased from the 25% to 75% quantile (blue to red) while keeping GDP per capita at the mean level. The right panel displays the transport when increasing the GDP per capita covariate from the 25% to 75% quantile (blue to red) while keeping mobility at the mean level. The dashed line corresponds to the identity function. The farther away the transport is from the identity map, the larger the lag is between deaths and infections, on average.

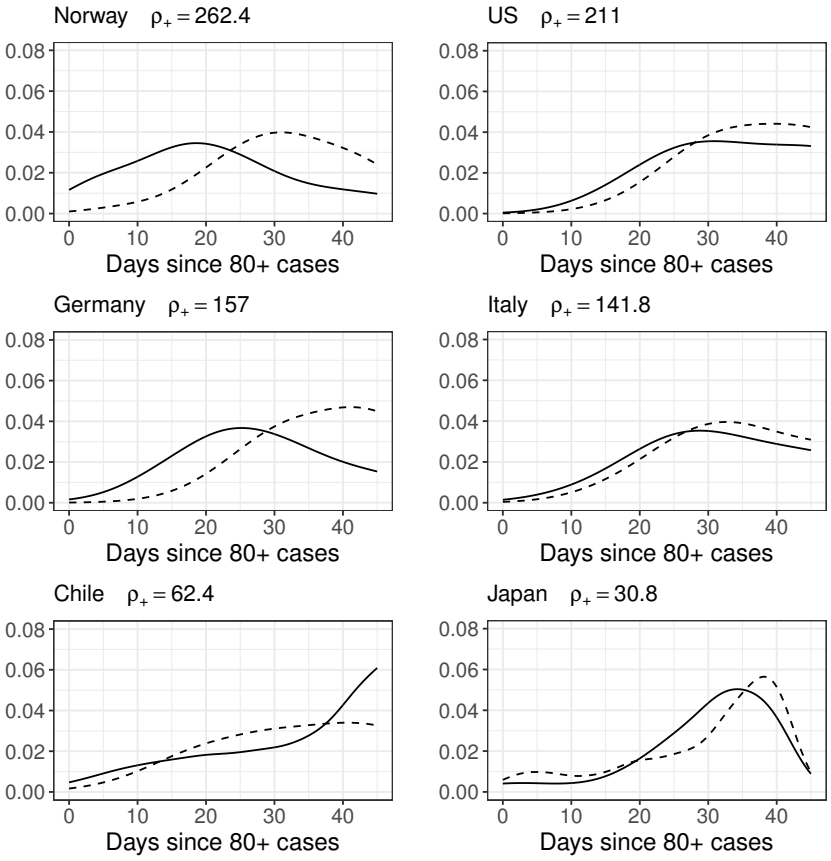

Figure 5. Estimated density functions for the arrival times of case (solid) and death (dashed) point processes for several countries, along with their corresponding $\rho_{\oplus}(x)$ values (see Methods 2.2).

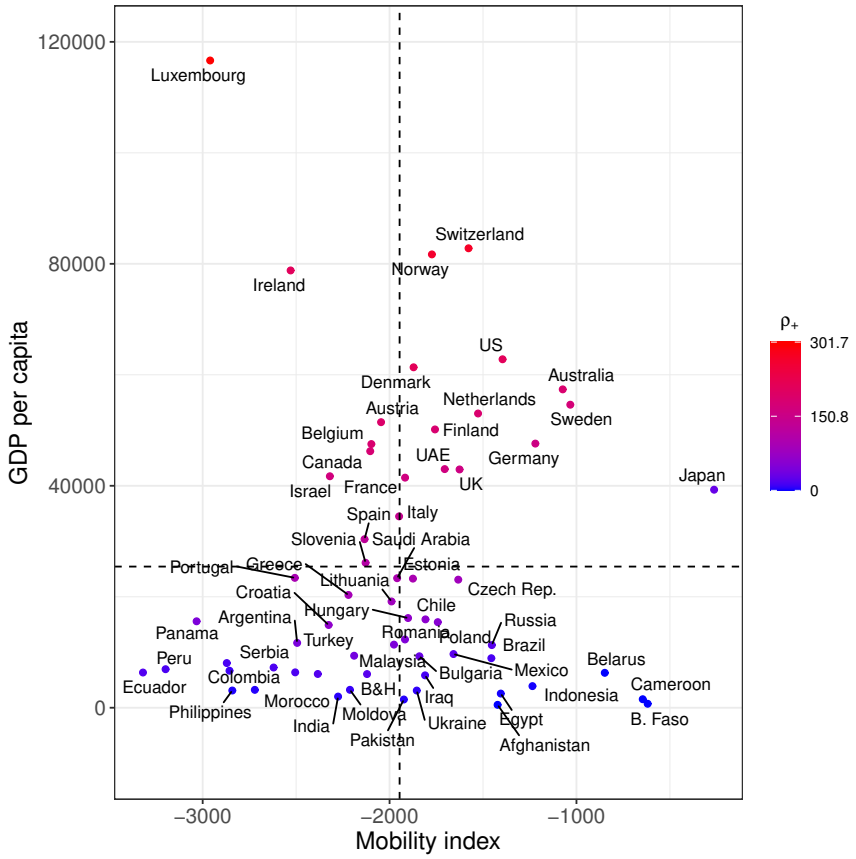

Figure 6. Positioning of the countries in terms of GDP per capita and Mobility index. Dashed lines correspond to the mean covariate levels. The country positions are colored based on the scalar $\rho_{\oplus}(x) = \int_0^T (T^{\oplus}_{c\to d}(t,x) - t) 1_{\{T^{\oplus}_{c\to d}(t,x)\geq t\}} dt$, which is a measure of the lag between deaths and infections, where this lag depends on the covariates (see Methods 2.4). Higher values indicate bigger time lags, i.e., deaths lag further behind cases. The first and third quartiles for the mobility index are -2309 and -1627, respectively, while 6352 and 42637 USD per person for the GDP per capita covariate.

However, there are exceptions. Japan provides an atypical example of a higher than average GDP per capita country with a low $\rho_{\oplus}(x)$ value, as opposed to other countries at similar wealth levels that have larger lags. This may be explained by its very high mobility index (more opportunities for infection) that might be associated with lower $\rho_{\oplus}(x)$ values (see Figure 4). In fact, the estimated fitted surface for $\rho_{\oplus}(x)$ on the covariates (see Figure **??** in the supplement) is clearly seen to decrease for sufficiently large values of the mobility index while holding the GDP per capita fixed, and moreover the decay is more pronounced for wealthier countries. Figure 5 shows the varying values for some countries. Norway's estimated arrival densities for the case and death process appear to be strongly time-warped and pushed apart, reflecting its high $\rho_{\oplus}$ value, while for countries with lower $\rho_{\oplus}$ values, such as Chile or Italy, the densities are much closer together.

4. Discussion

The replicated point process regression framework to analyze the COVID-19 cases and deaths while conditioning on covariates is well suited to lead to new insights and provide adequate modeling as it naturally takes into account the monotonicity and integer-valued property of the underlying case and death process curves. While the point process approach is useful for modeling the time-evolution of COVID-19 cases and deaths, the conclusions nevertheless need to be taken with a grain of salt as the data on both cases and deaths are subject to under-reporting and other distortions that may be country-dependent.

The point process framework allows to analyze the relation of both the time occurrence and the number of infection or death events in the presence of covariates, and a detailed analysis of the time lag of COVID-19 deaths in relation to the reported cases. We find that both increasing GDP per capita and increasing mobility are associated with a higher number of cases and deaths over the entire time window $[0, 45]$ days, with peaks around 30 and 35 days after reaching 80+ cases, respectively. Increasing levels of GDP per capita are related to increasing timing effects in opposite directions between the death and case arrival times: The death curve is pushed towards the end of the time window while the opposite effect occurs for the cases.

We also see indications that increments in either GDP per capita or mobility index, keeping the other covariate at the mean level, are related to shorter doubling times for both case and deaths processes, and thus worse dynamics of disease spread. Increases in the mobility index are seen to be associated with a faster spread of the virus as more people are in contact with each others, which ultimately also leads to an increase in deaths.

Funding

Research supported by NSF Grant DMS-2014626.

References

[1] C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, *Data-based analysis, modelling and forecasting of the covid-19 outbreak*, PloS one 15 (2020), p. e0230405.

[2] J. Bigot, *Statistical data analysis in the Wasserstein space*, arXiv:1907.08417 (2019).

[3] B.M. Bolstad, R. Irizarry, M. Åstrand, and T. Speed, *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias*, Bioinformatics 19 (2003), pp. 185–193.

[4] F. Caramelo, N. Ferreira, and B. Oliveiros, *Estimation of risk factors for covid-19 mortality-preliminary results*, MedRxiv (2020).

[5] C. Carroll, S. Bhattacharjee, Y. Chen, P. Dubey, J. Fan, A. Gajardo, X. Zhou, H.G. Müller, and J.L. Wang, *Time dynamics of covid-19*, Scientific Reports 10:21040 (2020).

[6] C. Carroll, A. Gajardo, Y. Chen, X. Dai, J. Fan, P.Z. Hadjipantelis, K. Han, H. Ji, H.G. Müller, and J.L. Wang, *fdapace: Functional Data Analysis and Empirical Dynamics* (2020). Available at `https://CRAN.R-project.org/package=fdapace`, R package version 0.5.3.

[7] P.E. Castro, W.H. Lawton, and E.A. Sylvestre, *Principal modes of variation for processes with continuous sample curves*, Technometrics 28 (1986), pp. 329–337.

[8] Y. Chen, A. Gajardo, J. Fan, Q. Zhong, P. Dubey, S. Bhattacharjee, K. Han, and H.G.

Müller, *frechet: Statistical Analysis for Random Objects and Non-Euclidean Data* (2020). Available at `https://github.com/functionaldata/tFrechet`, R package version 0.1.0.

[9] D.R. Cox and V. Isham, *Point Processes*, Chapman & Hall, London, 1980, Monographs on Applied Probability and Statistics. MR MR598033 (82j:60091)

[10] D.J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes: volume I: Elementary Theory and Methods, Second Edition*, Springer, New York, 2003.

[11] S. Engle, J. Stromme, and A. Zhou, *Staying at home: mobility effects of covid-19*, Available at SSRN (2020).

[12] J. Fan and I. Gijbels, *Local Polynomial Modelling and its Applications*, Chapman & Hall, London, 1996. MR MR1383587 (97f:62063)

[13] Á. Gajardo and H.G. Müller, *Point process regression*, arXiv preprint arXiv:2006.00447 (2020).

[14] D. Gervini, *Independent component models for replicated point processes*, Spatial Statistics 18 (2016), pp. 474 – 488.

[15] D. Gervini and M. Khanal, *Exploring patterns of demand in bike sharing systems via replicated point process models*, Journal of the Royal Statistical Society Series C: Applied Statistics 68 (2019), pp. 585–602.

[16] Google, *Google llc "google covid-19 community mobility reports". https://www.google.com/covid19/mobility/ accessed: May 24, 2020.* (2020).

[17] P. Hall, *The influence of rounding errors on some nonparametric estimators of a density and its derivatives*, SIAM Journal on Applied Mathematics 42 (1982), pp. 390–399.

[18] L. Horvath and P. Kokoszka, *Inference for Functional Data with Applications*, Springer, New York, 2012.

[19] JHU, *The covid-19 data repository by the center for systems science and engineering (csse) at Johns Hopkins University.* (2020).

[20] M.U. Kraemer, C.H. Yang, B. Gutierrez, C.H. Wu, B. Klein, D.M. Pigott, L. Du Plessis, N.R. Faria, R. Li, W.P. Hanage, *et al.*, *The effect of human mobility and control measures on the covid-19 epidemic in china*, Science 368 (2020), pp. 493–497.

[21] A.J. Kucharski, T.W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R.M. Eggo, F. Sun, M. Jit, J.D. Munday, *et al.*, *Early dynamics of transmission and control of covid-19: a mathematical modelling study*, The Lancet Infectious Diseases (2020).

[22] P.H. Lee, *Estimating the real-time case fatality rate of covid-19 using Poisson mixtures model*, medRxiv (2020).

[23] T. Liu, J. Hu, M. Kang, L. Lin, H. Zhong, J. Xiao, G. He, T. Song, Q. Huang, Z. Rong, A. Deng, W. Zeng, X. Tan, S. Zeng, Z. Zhu, J. Li, D. Wan, J. Lu, H. Deng, J. He, and W. Ma, *Transmission dynamics of 2019 novel coronavirus (2019-ncov)*, bioRxiv (2020).

[24] G. Mohler, F. Schoenberg, M.B. Short, and D. Sledge, *Analyzing the world-wide impact of public health interventions on the transmission dynamics of covid-19*, arXiv preprint arXiv:2004.01714 (2020).

[25] H.G. Müller and U. Stadtmüller, *Multivariate boundary kernels and a continuous least squares principle*, Journal of the Royal Statistical Society B 61 (1999), pp. 439–458.

[26] V.M. Panaretos and Y. Zemel, *Amplitude and phase variation of point processes*, The Annals of Statistics 44 (2016), pp. 771–812.

[27] V.M. Panaretos and Y. Zemel, *Statistical aspects of Wasserstein distances*, Annual Review of Statistics and its Application 6 (2019), pp. 405–431.

[28] L. Pellis, F. Scarabel, H.B. Stage, C.E. Overton, L.H.K. Chappell, K.A. Lythgoe, E. Fearon, E. Bennett, J. Curran-Sebastian, R. Das, *et al.*, *Challenges in control of covid-19: short doubling time and long delay to effect of interventions*, arXiv preprint arXiv:2004.00117 (2020).

[29] A. Petersen and H.G. Müller, *Functional data analysis for density functions by transformation to a Hilbert space*, Annals of Statistics 44 (2016), pp. 183–218.

[30] A. Petersen and H.G. Müller, *Fréchet regression for random objects with Euclidean predictors*, The Annals of Statistics 47 (2019), pp. 691–719.

[31] F. Petropoulos and S. Makridakis, *Forecasting the novel coronavirus covid-19*, PloS one

15 (2020), p. e0231236.
[32] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, 2003.
[33] J.L. Wang, J.M. Chiou, and H.G. Müller, *Functional data analysis*, Annual Review of Statistics and Its Application 3 (2016), pp. 257–295.
[34] K. Wu, D. Darcet, Q. Wang, and D. Sornette, *Generalized logistic growth modeling of the covid-19 outbreak in 29 provinces in china and in the rest of the world*, arXiv preprint arXiv:2003.05681 (2020).
[35] T. Zhang and G. Lin, *Spatiotemporal analysis for the outbreak of covid-19 in the world*, Available at SSRN 3576816 (2020).
[36] Z. Zhang and H.G. Müller, *Functional density synchronization*, Computational Statistics and Data Analysis 55 (2011), pp. 2234 – 2249.