# Estimating Causal Effects Using Gaussian Process

December 22, 2021

## 1 Introduction

In recent times, many application have been devoted to understanding and revealing *causal* rather than associative relations among variables. The two main building blocks of the literature pivot around Rubin Causal Model (Holland, 1986), a framework based on potential outcomes or counterfactuals, and causal graphs or structural model as in Pearl (1995). The former approach, which is going to be the one explored by this article, is based on the idea of recovering the outcome that could have been manifested if a specific event wouldn't have taken place.

This article contribution focus primarily on expanding and generalizing this class of model, allowing for non-linearity in a non-parametric manner through Gaussian Processes. This way, data can grant high degree of flexibility in building the counterfactual outcome, using all type of information one can gather without any limitations on the functional form. The proposed models make also possible to asses the robustness of the synthetic controls, as we can use posterior distribution of the Gaussian Processes to quantify uncertainty. Lastly, as the model learn the relationship which prevail among variables, there is no need to match time series start date and end date, making the most out of available data.

To our best knowledge the only paper that uses Gaussian process in the context of potential outcomes is Alaa and van der Schaar (2017). The purpose of the article was to infer individualized treatment effect across a series of cross sectional experiments. Nonetheless, the bivariate settings arise from the use of the treated and control group as dependent variable and no time component is exploited. On the other hand, there exist a very recent literature that explores multi task causal learning using Gaussian process exploiting Judea Pearl's Do-Calculus (Pearl, 1995). These papers (Aglietti et al., 2020) however are mainly focused on understanding the main correlation structure of multiple continuous intervention functions - defined with a directed acyclic graph (DAG) - as opposed to a

single discrete intervention. Furthermore, the main data domain consist of cross sectional experiments, not time series data.

This paper is structured as follows. In sections 2 we introduce the causal framework and the synthetic control approach, presenting our main assumptions. In Section 3 we define Gaussian Process and the main model. In Section 4, we present the estimation procedure. In Section 5 we show an empirical application of our approach to describe the causal effect of UK effective vaccination program against other countries. Finally, the last section display our concluding remarks.

# 2 Background

## 2.1 Causal Framework

In this subsection, we build the framework to estimate the causal effect of an intervention on the treated subject. Each subject $y_{i,t}$ is associated with a $d$-dimensional feature vector $\mathbf{x}_{i,t} \in \mathcal{X}$ and a binary potential outcomes $y_i^{(w_i)} \in \mathbb{R}$ where $w_i \in \{0, 1\}$ is a treatment assignment indicator with "1" referring to the variable being treated and "0" as control. Throughout the paper we assume that the covariates are independent from the treatment i.e. $\mathbf{x}_{i,t}^{w_i} = \mathbf{x}_{i,t}^{1-w_i}$. Let $\delta_{i,t} = y_{i,t}^{(1)} - y_{i,t}^{(0)}$, then the *additive causal* effect on the subject $i$ is the population average treatment effect and it is given by

$$\tau_{i,t} = \mathbb{E}\left(\delta_{i,t} | \mathbf{x}_{i,t}\right). \tag{1}$$

We are also interested in the uncertainty surrounding the treatment effect. This can be either be measured through the variance

$$\sigma_{\delta,t}^2 = \mathbb{V}\left(\delta_{i,t} | \mathbf{x}_{i,t}\right), \tag{2}$$

or directly applying quantile function to calculate credible regions

$$q_t^{\delta}(\alpha) = F_{\alpha}^{-1}\left(\delta_{i,t} | \mathbf{x}_{i,t}\right), \tag{3}$$

with level of confidence generally $\alpha = 5\%$ and $\alpha = 95\%$. We aim to estimate those measures from a dataset $\mathcal{D} = \{X, \mathbf{y}, \boldsymbol{w}\}$, which involve T samples of different time series. The main challenge lays on the fact that we only observe one of the potential outcomes for every subject $i$, i.e. the treatment effect is unobserved, so we can't use a general supervised learning tool to estimate $\tau_t$. In addition to its point-wise impact, we could be

interested in the over-time cumulative effect of the intervention

$$\mathcal{T}_i = \sum_{t=t_0+1}^{T} \tau_{i,t} \qquad \forall t = t_0 + 1, ...T, \tag{4}$$

where $t_0$ represent the time in which the intervention take place. The cumulative sum is a valuable measure when $\mathbf{y}$ is a *flow* variable which is measured over an interval of time (e.g number of death in a country). This quantity however lose its interpretability when $\mathbf{y}$ is a *stock* variable, a quantity measured at a specific time, and represents a quantity existing at that point in time (e.g rate of infectiousness). In this case is more appropriate to use the average treatment effect after the intervention

$$\overline{\tau}_i = \frac{1}{T - t_0} \sum_{t=t_0+1}^{T} \tau_{i,t} \qquad \forall t = t_0 + 1, ...T, \tag{5}$$

This measure extends to the time series framework of the average distributional shift effect in (Sävje et al., 2019), with the difference that here it is averaged across time as opposed to units.

Within a Gaussian framework, expected value and variances are straightforward to derive and we have that $\delta_t \sim \mathcal{N}(\tau_t, \sigma^2_\tau)$. Sometimes however, Gaussian likelihood do not fit data properly and some mathematical transformation may be needed. For, example a random variable which takes only non-negative value (e.g. counting, lengths) would be better represented using a log-normal distribution. Thus, if a random variable $\delta_t = \log(\delta^*_t)$ has a Gaussian distribution, then $\delta^*_t \sim \log \mathcal{N}(\tau^*_t, \sigma^2_{\delta^*})$ is log-normally distributed. By modelling directly the transformed variable one obtain that the causal effect given by

$$\delta^*_{i,t} \;=\; \exp\left(\log y^{(1)}_{i,t} - \log y^{(0)}_{i,t}\right) = \frac{y^{(1)}_{i,t}}{y^{(0)}_{i,t}} \tag{6}$$

Then, taking expectation $\mathbb{E}(\delta^*_t|\mathbf{x}_{i,t}) = \tau^*_t$ and using the fact that $\delta^*_t$ is log-normal

$$\tau^*_{i,t} \;=\; \mathbb{E}\left(\frac{y^{(1)}_{i,t}}{y^{(0)}_{i,t}}\Big|\mathbf{x}_{i,t}\right) = \exp\left(\tau_t + \frac{\sigma^2_{\delta,t}}{2}\right), \tag{7}$$

with related percentiles as in 3. This can be interpreted as a *multiplicative causal effect* with base 1. A ratio bigger than 1 will indicate a positive impact on the treated individual, while if smaller, it can indicate a negative effect.

Generally for the cumulative effect (4) and average effect (5), there is no closed form solution unless each $\delta_t$ is normally distributed and independent over time. In this case one can use samples from the posterior predictive distribution over the counterfactual variable to obtain samples from the posterior causal effect distribution, the quantity we are interested in. This methods work as well when using variable transformations as one can convert it back to the original scale and then calculate the empirical cumulative (average) distribution, with given mean and quantiles.

## 2.2 Potential Outcomes - Synthetic Control

Synthetic control methods have gained traction as method to estimate causal effect from variables that were subject to a single intervention or treatment. A traditional approach is the one based on the *difference-in-difference* (DD) method, a static linear regression model where the causal effect is estimated as the difference between the regression coefficient in the treated group and the one in the control group. This is often implemented in a linear regression setting, with the quantity of interest being the interaction term of the proxy variable and the treatment group dummy variable. However, DD methods suffer from two main drowbacks Brodersen et al. (2014): the first one is that it assumes that the data have an i.i.d. distribution, thus disregarding the temporal component; secondly, only two time points are effectively considered: the pre and post intervention period. Partial improvement of these models has recently been offered by (Abadie et al., 2010) (Abadie and Gardeazabal, 2003). The proposed models generalize the DD as they allow the effect of unit-specific unobserved variables to vary with time. In particular, they recover the counter-factual outcome using a control group that has a similar pattern in the pre-intervention period as the treated unit. To do so they find a matrix of weight which minimize the squared distance between the pre-intervention features of the exposed region and the the features for the unaffected regions. However, this method has its own limitation. Indeed, it focuses only on possible convex combinations of control time series to match the treated variable. Furthermore, there is a non-negligible data loss in regards to the temporal component. First of all, only data in the pre-treatment period is used to construct the counter-factual unit. Secondly, time series evolution and interaction over time is neglected, as data is aggregated over time or treated individually for each time period. An alternative class of models is identified by Brodersen et al. (2014), whose approach addresses many of the previous methods limitations. The authors approach relies on Bayesian state space models, which encompass the outcome's temporal evolution with exogenous regression components to efficiently build a counter-factual model. State space allow for flexibility when modelling a variable that could be exposed by external

noise, distinguishing between a state equation which describe the transition of the latent variable from one point in time to the next one, and a measurement equation, which describe the accuracy of the signal. Additionally, being fully Bayesian make it possible to (i) incorporate prior probability about the model structures and parameters and (ii) have a posterior distribution, and thus a probability interpretation, for the causal impact of the intervention. Although the models focus on one outcome variable and multiple controls, an extension to a multivariate setting has been implemented using Multivariate Bayesian Structural Time Series (Menchetti and Bojinov, 2020), which is limited remain limited to linear relationships between outcomes and controls and subject to the Markovian assumption of the variables.

# 3 Multi Output Gaussian Process

Gaussian Processes (GPs) are a powerful Bayesian method for regression and classification problems. GPs take into account non-linear dependencies between inputs and outputs by making the covariance between two points take the form of non-linear functions (Rasmussen and Williams, 2006). Given their Bayesian nature, GPs also provide a natural way to construct and handle uncertainties. A straightforward extension of these models are the Multi-output Gaussian processes (MOGP), which generalise GPs in a multivariate framework. MOPG exploits correlations between multiple outputs and across the input space, providing better predictions, particularly in scenarios with noisy data or missing values (Bonilla et al., 2008). In this paper, we are going to focus on a class of model referred to as Linear Model of Coregionalization (LMC), in which each output corresponds to a linear combination of one or more latent random function.

## 3.1 The Model

Define $\mathbf{y} = \{\mathbf{y}_1^{'}, \cdots, \mathbf{y}_m^{'}\}^{'}$ where $\mathbf{y}_j \in \mathbb{R}^{T_j}$ as the vector of observed variables and $X = \{X_1^{'}, \cdots, X_m^{'}\}^{'}$with $X_j \in \mathbb{R}^{T_j \times d}$ the matrix of the $d$ covariates associated with output $j$. For the training set we assume an *heterotopic data* configuration, i.e. each output have different training sets $X_1 \neq, \cdots, \neq X_m$ each one with $T_j$ samples such that $T = \sum_{j=1}^{m} T_j$ (Liu et al., 2018). A Gaussian processes model is shown below,

$$\mathbf{y}_j(X_j) = \boldsymbol{\tau}_j + f_j(X_j) + \boldsymbol{\epsilon}_j, \;\; \boldsymbol{\epsilon}_j \sim \mathcal{N}(0, \omega_j^2 \mathrm{I}_{T_j}), \tag{8}$$

for each $j = 1, \cdots, m$ and where where the iid noise $\epsilon$ accounts for the observation errors. The likelihood function for the $m$ outputs is defined as

$$p(\mathbf{y} - \boldsymbol{\tau} | \mathbf{f}(X), X, \Omega) = \mathcal{N}(\mathbf{f}(X), \Omega), \tag{9}$$

where $\Omega = \text{diag}(\omega_1^2 \mathrm{I}_{T_1}, \cdots, \omega_m^2 \mathrm{I}_{T_m}) \in \mathbb{R}^{T \times T}$ and the outputs $\mathbf{f}(X) = \{f_1(X_1), \cdots f_m(X_m)\}'$ are probability distributions in function space and represent the Multi Output Gaussian Process (MOPG)

$$\mathbf{f}(X) \sim \mathcal{GP}(\mu(X), \mathcal{K}(X, X)). \tag{10}$$

Without loss of generality one can assume that $\mu(X) = \mathbf{0}$ while $\mathcal{K}(X, X) \in \mathbb{R}^{mT \times mT}$ is the multi-output positive semidefinite covariance matrix, defined as

$$\mathcal{K}(X, X) = \begin{bmatrix} K_{1,1}(X_1, X_1) & K_{1,2}(X_1, X_2) & \dots & K_{1,m}(X_1, X_m) \\ K_{2,1}(X_2, X_1) & K_{2,2}(X_2, X_2) & \dots & K_{2,m}(X_2, X_m) \\ \vdots & \vdots & \ddots & \vdots \\ K_{m,1}(X_m, X_1) & K_{m,2}(X_m, X_2) & \dots & K_{m,m}(X_m, X_m) \end{bmatrix} \tag{11}$$

with $K_{i,j}(X_i, X_j) = K_{j,i}(X_j, X_i)' \ \forall i, j$ by symmetry. Taking a look at block matrices $K_{i,j}(X_i, X_j) \in \mathbb{R}^{T_i \times T_j}$, they are defined such that

$$K_{i,j}(X_i, X_j) = \begin{bmatrix} k_\theta(\mathbf{x}_{i,1}, \mathbf{x}_{j,1}) & \dots & k_\theta(\mathbf{x}_{i,1}, \mathbf{x}_{j,T_j}) \\ \vdots & \ddots & \vdots \\ k_\theta(\mathbf{x}_{i,T_i}, \mathbf{x}_{j,1}) & \dots & k_\theta(\mathbf{x}_{i,T_i}, \mathbf{x}_{j,T_j}) \end{bmatrix}, \quad i, j = 1, \dots, m.$$

The next step is to define the kernel for the covariance of each of the GPs $f(\mathbf{x}_j)$. For simplicity, let us focus first on the case of $i = j$, so we can drop the unit subscript.
The squared exponential kernel is a popular choice:

$$k_\theta(\mathbf{x}_s, \mathbf{x}_t) = \sigma^2 \exp\left(-\frac{\|\mathbf{x}_s - \mathbf{x}_t\|^2}{2\ell^2}\right).$$

with $\boldsymbol{\theta} = (\ell, \sigma^2)'$. Another useful kernel is the Matérn kernel given by:

$$k_\theta(\mathbf{x}_s, \mathbf{x}_t) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} \|\mathbf{x}_s - \mathbf{x}_t\|^2\right)^\nu B_\nu \left(\frac{\sqrt{2\nu}}{\ell} \|\mathbf{x}_s - \mathbf{x}_t\|^2\right) \tag{12}$$

with $\boldsymbol{\theta} = (\ell, \sigma^2)'$ and where $B_\nu(\cdot)$ is the modified Bessel function and $\Gamma(\cdot)$ is the gamma function. When the dimension $d = 1$ we have that this kernel generate a continuous-time version of an AR(p) Gaussian process such that $p = \nu - 1/2$. A particular case is achieved with $\nu = 1/2$, since the Matérn kernel reduces to the exponential kernel given by $k_\theta(\mathbf{x}_s, \mathbf{x}_t) = \exp(\|\mathbf{x}_s - \mathbf{x}_t\| / \ell)$ which is the covariance process of a Ornstein-Uhlenbeck (OU) process, the continuous-time analogue of an AR(1) process. Let us focus on the one-dimensional case with $\mathbf{x}_t = t$ and let us call the lag between two time point $h$. It is

shown that taking the Fourier transform of the power spectrum of an OU process on $\mathbb{R}$ with drift $\phi$ and diffusion $\sigma$ gives

$$k(h) = \frac{\sigma^2}{2\phi} e^{-\phi|h|} \tag{13}$$

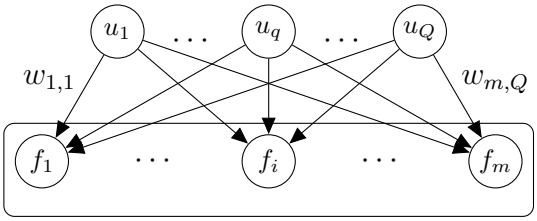
Thus, the exponential decay rate of the autocorrelation is captured using $t$ instead of a lagged version of $y_t$. Both the above kernels are stationary, i.e. the covariance function depends on the relative positions of two inputs and not their absolute location.[1] If $k(x_1, x_1)$ and $k(x_2, x_2)$ are covariance functions over different spaces $\mathcal{X}_1$ and $\mathcal{X}_2$, then the direct sum $k(x, x) = k_1(x_1, x_1) + k_2(x_2, x_2)$ and the tensor product $k(x, x) = k_1(x_1, x_1) \cdot k_2(x_2, x_2)$ are also covariance functions (defined on the product space $\mathcal{X}_1 \times \mathcal{X}_2$), by virtue of the sum and product constructions. We can then flexibly sum or multiply all the kernels to have the type of interaction we need in the covariance matrix. As an example a linear kernel plus a periodic one will generate a periodic kernel with a trend.

Finally, in order to fully specify the distribution of $\mathbf{f}(X)$, being a GP with multiple outputs, we need to make an assumption on the dependence between $f_j(\mathbf{x}_j)$'s. The simplest way to go is to assume independence which will imply the following covariance structure $\mathcal{K} = \text{ diag}(K_1(X_1, X_1), ..., K_m(X_m, X_m))$. There are more flexible dependence structures among the $f_j(\mathbf{x}_j)$'s and specifying the covariance of $\mathbf{f}(X)$, $\mathcal{K}(X, X)$; see for example Alvarez et (2012) for a survey of several methods including the intrinsic coregionalisation model (ICM), the semiparametric latent factor model (SLFM) and the linear model of coregionalisation (LMC). These models may be viewed as performing exploratory factor analysis on $\mathcal{K}$ with unobserved factors $u_q(X)$. In particular, let us consider the LMC. This specification, widely used in geostatistics, expresses the outputs as a linear combination of Q latent functions as

$$f_j(X_j) = \sum_{q=1}^{Q} w_{j,q} u_q(X_j) \tag{14}$$

where $u_q(X_j)$ is itself a latent Gaussian process with mean $\mathbf{0}$ and $\text{cov}[u_q(X_j), u_q(X_j)] = K_q(X_j, X_j)$, while $w_{j,q}$'s are the coefficients which measure output correlations. This is because the model introduce in independence assumption between $u_q(X_j)$ and $u_p(X_j)$ for $p \neq q$, such that $\text{cov}[u_q(X_j), u_p(X_j)] = 0$

[1] To allow some flexibility in the model, especially when data exhibit visible trends, one can introduce *non-stationary* kernels. The most simple example is the linear kernel. This is defined as $k_\theta(\mathbf{x}_s, \mathbf{x}_t) = \sigma_i^2 \mathbf{x}_s^{'} \mathbf{x}_t$ where $\theta = \sigma^2$




Then, cross covariances between the output can be calculated by

$$K_{j,i}(X_j, X_i) = \sum_{q=1}^{Q} w_{j,q} w_{i,q} K_q(X_j, X_i) \tag{15}$$

Linear combination of different kernels still result in a valid positive definite covariance matrix. This approach is defined separable (Alvarez et al., 2012) due to the decoupled input-output structure of the covariance.
Define $\mathrm{B}_q$ the rank 1, $m \times m$ positive semi-definite matrix such that $\mathrm{B}_q = \mathbf{w}_q \mathbf{w}_q^{'} + \mathrm{diag}(k_{1,q}, \cdots, k_{q,m})$ with $k_{j,q}$ positive and $\mathbf{w}_q = \{w_{1,q}, \cdots, w_{m,q}\}^{'}$. Then, one can write the multi-output covariance as

$$\mathcal{K}(X, X) = \sum_{q=1}^{Q} \mathrm{B}_q \otimes K_q(X, X) \tag{16}$$

The coregionalization matrix B defines the amount of inter and intra task transfer of learning among all the outputs. Thus, the latent kernel is shared across all the output but is scaled by a factor $\mathrm{B}_{[i,j]}$.[2] Denote by $\overline{X} \in \mathbb{R}^{T \times (d-1)}$ the matrix of $d$ inputs minus *time*, which is denoted by $t$. Let us consider specifically the time series defined in (8). Using the structure of 16 we will focus on a particular specification given by

$$\mathcal{K} = \mathrm{B}_1 \otimes K_{rbf}(\overline{X}, \overline{X}) + \mathrm{B}_2 \otimes K_{Mat}(t, t), \tag{17}$$

where $K_{rbf}(\cdot), K_{Mat}(\cdot)$ are the squared exponential and Matérn respectively. In this way one can capture both stationary trends between the output and the input and autocorrelation structure for each output (the latter given by continuous generalization of an AR($p$) process).
Typically, we employ training data to incorporate the knowledge about the function to find the posterior distribution for an arbitrary set of test location $\mathbf{x}_*$. GPs are stochastic

[2]The special case of $Q = 1$, generate to the intrinsic coregionalization model (ICM). The computational complexity is largely reduced in exchange of a more restrictive architecture, as one latent process becomes only the source of variability among outputs.

process in which any finite subset of random variables follows a joint normal distribution. Thus, it is possible to determine the joint prior distribution of the observations $\mathbf{y}$ and the output $\mathbf{f}_* = \mathbf{f}(X_*)$ at test points $X_*$ as

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left[ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathcal{K}(X,X) + \Omega & \mathcal{K}(X,X_*) \\ \mathcal{K}(X_*,X) & \mathcal{K}(X_*,X_*) \end{pmatrix} \right] \tag{18}$$

where $\mathcal{K}(X, X_*) \in \mathbb{R}^{T \times T_*}$ is matrix of the covariances calculated at all pairs of training and test points, $X$ and $X_*$. Then, it is possible to analytically derive the posterior distribution of $\mathbf{f}_*$, conditioned on $\mathbf{y}$, by using multivariate Gaussian proprieties.

$$\mathbf{f}_* | \mathbf{y}, X, X_* \sim \mathcal{N}(\hat{\mathbf{f}}_*, \Sigma_*) \tag{19}$$

where $\hat{\mathbf{f}}_*$ and $\Sigma_*$ are the predictive mean and predictive variance, given by

$$\hat{\mathbf{f}}_* = \mathcal{K}(X_*, X)[\mathcal{K}(X,X) + \Omega]^{-1}\mathbf{y} \tag{20}$$

$$\Sigma_* = \mathcal{K}(X_*, X_*) - \mathcal{K}(X_*, X)[\mathcal{K}(X,X) + \Omega]^{-1}\mathcal{K}(X, X_*) \tag{21}$$

Thus, the predictive uncertainty $\Sigma_*$, does not depend on $\mathbf{y}$, but only on the output dependencies given by the kernel structure of $X$ and $X_*$.
To complete the model one also needs to assign suitable priors on the parameters.

## 3.2 Methodology

To assess the ability of the trained model to generalize the data, we adopt a typical machine learning approach. In particular, we evaluate the performance of the trained model before the intervention ($t < t_0$). The new dataset is split into two part, the train and the validation set, which account for around 33% and 67% respectively.
Primarily, there are two main issues to address: which countries to choose as control series and which type of kernel structure the model should have.

For the first one, given the high number of combinations of countries possible we restrict the set by using Dynamic time warping, or DTW (Giorgino, 2009). The algorithm produces a distance metric between two input time series. The similarity or dissimilarity of two time-series is then calculated by converting the data into vectors and calculating the Euclidean distance between those points in vector space. The 10 components which minimise the distance are chosen to be the set of potential control series of the experiment. This algorithm is particularly useful for dealing with sequences in which single components have characteristics that vary over time, not necessarily in sync.

The second issue is related to the appropriate choice of the LMC architecture and relevant kernel function. We compare different methods:

1) *Semiparametric Latent Factor GP* (SLFGP). This is the model outlined in the section 3.1 equation (17), in which the rbf kernel is adopted on the input space given by the spatial covariates and the Matèrn Kernel on the time trend.

2) *Independent GPs* (INGP). While the coregionalized model shares information across outputs, the independent models cannot do that. In particular we assume that in (16), $\mathbf{w}_q = \mathbf{0}$ and $k_{j,q} = 1 \ \forall \ j, q$, i.e. $\mathrm{B}_q = \mathrm{I}_m$. In the regions where there is no training data specific to an output the independent models tend to revert to the prior assumptions.

3) *One Factor GP* (1FGP). Instead of assuming two separate input space and kernels for time and the other covariates as in (16), we combine the two Kernel such that $\mathcal{K} = \mathrm{B}_1 \otimes (K_{rbf}(X, X) + K_{Mat}(X, X))$. In this way we combine feature from both $rbf$ and *Matérn* on a shared input space.

4) *Two RBF Factor GP* (2RBFGP). The input space is divided into time trend and spatial covariates but the kernel function has the same structure (Radial Basis Function) for both $t$ and $X$. $\mathcal{K} = \mathrm{B}_1 \otimes K_{rbf}(\overline{X}, \overline{X}) + \mathrm{B}_2 \otimes K_{rbf}(t, t)$.

5) *Bayesian Causal Impact* (BCI). The model outlined by (Brodersen et al., 2014), in which the outcome variable is $\mathbf{y}_i$ and the covariates are $(\{\mathbf{y}_j\}_{j \neq i}, X_i)$, i.e. all the other control variables and the relevant covariates for the treated subject $i$. In this case we adopt an *isotopic data* framework. [3]

Once the models have been fitted on the train dataset we can evaluate the forecast performance calculating the distance from the ground truth. We use different measure of dispersion

1) *Mean Squared Error* (MSE). Being the most simple form of dispersion, it computes the average of the squared errors, calculated as the differences between the estimated values and the actual values

$$\mathrm{MSE}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} (y_{i,t} - \hat{f}_{i,t}(\mathbf{x}^*_{i,t}))^2. \tag{22}$$

[3] For the Bayesian Model of (Brodersen et al., 2014) one can use only the number of point such that $T_i = \min(T_1, ...T_m)$ while for the Gaussian Process we have heterotopic data in which $T_i$ may be different from $T_j$.

Still, it measure point estimates disregarding the uncertainty of the prediction.

2) *Log Score* (LogS). Forecasts are usually surrounded by uncertainty, and being able to quantify it is pivotal to good decision making. The logarithmic score (Good, 1952), defined as

$$\text{LogS}_i = -\frac{1}{T_i}\sum_{t=1}^{T_i} \log \mathcal{N}(\hat{f}_{i,t}(\mathbf{x}^*_{i,t}), \sigma_{i,t}(\mathbf{x}^*_{i,t})) \tag{23}$$

and is equal to the log of the predictive density of $\mathbf{y}_i$ given by (19). The measure take into consideration also the variability of the point forecast. Since we are working with GP the distribution is Normal and available in closed form, making the calculation straightforward.

3) *Energy Score* (ES). This scoring rule is the multivariate extension of the the continuous ranked probability score, CRSP (Matheson and Winkler, 1976). Let $\mathbf{y} = \{y_1, \cdots, y_h\}' \in \mathbb{R}^h$ the values of the outcome $i$ on the $h$-horizon test set. Denote $\mathbf{f}_*$ the forecast distribution on $\mathbb{R}^h$ with N samples $\{\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(N)}\}$ with $\mathbf{z}^{(k)} = \{z_1^{(k)}, \cdots, z_h^{(k)}\}$ from $\mathbf{f}_*$ with $k = 1, \cdots, N$, Then, the energy score can be calculated as

$$\text{ES}_i = \frac{1}{N}\sum_{k=1}^{N} ||\mathbf{z}^{(k)} - \mathbf{y}|| - \frac{1}{2N^2}\sum_{k=1}^{N}\sum_{c=1}^{N} ||\mathbf{z}^{(k)} - \mathbf{z}^{(c)}|| \tag{24}$$

where $||\cdot||$ is the Euclidean norm on $\mathbb{R}^h$. This function evaluates samples from a multivariate forecast an return a single estimate. [4]

For a given model, the lower the score, the highest the accuracy of the forecast. All three measure are compared to see which model is better.

# 4 Estimation

The Bayesian framework provides effective and consistent tools for inference. As for many applications in machine learning, the advanced computations coming from the integration over the parameter space are usually analytically intractable, and effective approximation methods require complex derivations. GPs, however, can be treated as hierarchical models, where the parameters are represented by the latent function $\mathbf{f}(X) = \mathbf{f}$, which in

[4]Sampling from the forecast distribution can be regarded as an approximation the values of the proper scoring rules, for a sufficient large N (Jordan et al., 2019)

turn can be considered samples from a population characterised by hyper-parameters $\theta$'s. In this case, the $\theta$'s are the parameter of the kernel covariance functions and likelihood variances $\omega_i^2$. Given Bayes rule, the posterior over the parameters is

$$p(\mathbf{f}|\mathbf{y}, X, \boldsymbol{\theta}) = \frac{p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|\boldsymbol{\theta})}{\int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|\boldsymbol{\theta})\, d\mathbf{f}} \tag{25}$$

where $p(\mathbf{y}|\mathbf{f}, X)$ is the *likelihood*, $p(\mathbf{f}|\boldsymbol{\theta})$ is the prior and the expression in the denominator is a normalizing constant, called *marginal likelihood*. Then, we can then express the hyperparameters posterior, making the marginal likelihood from above play the role of the likelihood so that

$$p(\boldsymbol{\theta}|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{y}|X, \boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \tag{26}$$

In practice, instead of maximizing the posterior in (26), one can instead maximize the marginal likelihood, with respect to the hyperparameters $\boldsymbol{\theta}$ (*Type II Maximum Likelihood*)

$$p(\mathbf{y}|X, \boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f}, X, \theta)p(\mathbf{f}|X, \boldsymbol{\theta})\, d\mathbf{f} \tag{27}$$

The strength of GPs lays on the tractability of the integral over the parameters $\mathbf{f}$, since we know that $p(\mathbf{f}|X, \theta) = \mathcal{N}(\mathbf{f}|0, \mathcal{K})$. Furthermore, we have that $p(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \Omega)$. Then, following Rasmussen and Williams (2006), one can perform the integration of the product of two normal which which yield the log marginal likelihood

$$\log p(\mathbf{y}|X, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}'\Sigma\mathbf{y} - \frac{1}{2}\log|\Sigma| - \frac{T}{2}\log(2\pi), \tag{28}$$

where $\Sigma = \mathcal{K} + \Omega$ is the covariance matrix of the noisy outcome $\mathbf{y}$ and contains all *hyper-parameters*. The first term is a data-fit terms, as it is the only one involving $y$, the second one represent the complexity term, since it depends only on the covariance function, and the last one is just a constant. Marginalising out the Gaussian vector f, moves up the Bayesian hierarchy by one level, thus reducing the odds of overfitting (Murphy, 2013). To maximize the marginal likelihood we first find the derivative of the marginal likelihood with respect to the kernel hyperparameters

$$\frac{\partial}{\partial\theta_i}\log p(\mathbf{y}|X, \boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}'\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}\mathbf{y}' - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right), \tag{29}$$

where $\frac{\partial\Sigma}{\partial\theta_i}$ depends on the structure of the kernel and the parameters we are taking derivatives of. The inversion of the $\mathcal{K}$ matrix requires $\mathcal{O}(n^3)$ by standard method, and then

$\mathcal{O}(n^2)$ time per hyperparameters to calculate the gradient. Given the minor relative computational cost of calculating derivatives, a gradient based optimizer would be beneficial.[5] A very popular method is BFGS, named after its inventors Broyden, Fletcher, Goldfarb and Shanno. As a Quasi-Newton procedure it approximates the Hessian using the differences of gradients over several iterations, thanks to a *secant* (Quasi-Newton) condition.

**Algorithm 1** BFGS method

choose initial guess $\boldsymbol{\theta}_0$
choose$B_0$, the initial Hessian guess, e.g. $B_0 = \mathrm{I}$
**for** $k = 0, 1, 2, \cdots$ **do**
    solve $B_k \boldsymbol{s}_k = -\nabla f(\boldsymbol{\theta}_k)$
    $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \boldsymbol{s}_k$
    $\mathbf{y}_k = \nabla f(\boldsymbol{\theta}_{k+1}) - \nabla f(\boldsymbol{\theta}_k)$
    $B_{k+1}^{-1} = B_k + \frac{\mathbf{y}_k \mathbf{y}_k^{'}}{\mathbf{y}_k^{'} \boldsymbol{s}_k} - \frac{B_k \boldsymbol{s}_k \boldsymbol{s}_k^{'} B_k}{\boldsymbol{s}_k^{'} B_k \boldsymbol{s}_k}$
**end for**

The standard BFGS method employs the full history of gradients to calculate the Hessian approximation. On the other hand, the limited memory BFGS, abbreviated as L-BFGS, is based only on the most recent s (usually 20) gradients to compute the product $B_k^{-1} \nabla f(\boldsymbol{\theta}_k)$. The main advantage of L-BFGS is that is requires only a smaller storage requirement than $n(n+1)/2$ elements required to fully store the Hessian estimate, requiring only $\mathcal{O}(sn)$ instead of $\mathcal{O}(n^2)$ (Nocedal and Wright, 2006).
The L-BFGS-B algorithm further extends L-BFGS to handle linear constrain on variables such that $l_i \leq \theta_i \leq u_i$ where $l_i$ and $u_i$ are constant lower and upper bounds for each $\theta_i$. The algorithm separates fixed and unconstrained variables at each step by using the gradient method. Subsequently, it employs the L-BFGS method on the free variables to achieve higher accuracy.

Bayesian alternatives do exist but the most effective methods rely on MCMC, which can be quite slow for high. It is possible to approximate the posterior over the latent functions and over the hyper-parameters after setting the priors, using Hamiltonian Monte Carlo (HMC). Here, an additional *momentum* variable $\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \mathrm{M})$ is introduced for each parameter $\boldsymbol{\theta}$, the latter being regarded as *position*. The covariance matrix M, called

[5] Generally, the objective function is non-convex and local minima exist and can make the the optimization procedure challenging. However, empirical studies with non-complex covariance functions seem to indicate that the issue is not extremely serious, as every local maxima correspond to a different interpretation of the data (Rasmussen and Williams, 2006)

the mass matrix, rotate and scale the target distribution and it is generally set to the identity matrix, M = I, when no information is available on the target distribution. The joint density $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ defines the Hamiltonian

$$\begin{aligned} H(\boldsymbol{\phi}, \boldsymbol{\theta}) &= -\log p(\boldsymbol{\phi}, \boldsymbol{\theta}) && (30) \\ &= -\log p(\boldsymbol{\phi}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) && (31) \\ &= T(\boldsymbol{\phi}|\boldsymbol{\theta}) + V(\boldsymbol{\theta}). && (32) \end{aligned}$$

The first term $T(\boldsymbol{\phi}|\boldsymbol{\theta}) = -\log p(\boldsymbol{\phi}|\boldsymbol{\theta})$ is called "kinetic energy" and it is equal to the squared of the momentum since $-\log p(\boldsymbol{\phi}|\boldsymbol{\theta}) = \log p(\boldsymbol{\phi}) = 0.5\boldsymbol{\phi}'\boldsymbol{\phi}$, being he momentum density independent of the target density. The second term $V(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta})$ is the "potential energy" and is related to the target distribution $p(\boldsymbol{\theta})$. This extended model then follows Hamiltonian dynamics through fictitious time, whose evolution depend on a set of differential equation

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{dt} &= +\frac{\partial H}{\partial \boldsymbol{\phi}} = \frac{\partial T}{\partial \boldsymbol{\phi}} && (33) \\ \frac{d\boldsymbol{\phi}}{dt} &= -\frac{\partial H}{\partial \boldsymbol{\theta}} = -\frac{\partial T}{\partial \boldsymbol{\theta}} - \frac{\partial V}{\partial \boldsymbol{\theta}} = -\frac{\partial V}{\partial \boldsymbol{\theta}}, && (34) \end{aligned}$$

since $\frac{\partial T}{\partial \boldsymbol{\theta}} = 0$ by independence. The solution to these differential equation is not available in closed form and must be computed numerically. The most popular numerical integrators, which preserve volume and reversibility of the system is the *Leaprfog* integrator (Girolami and Calderhead, 2011). The leapfrog integrator takes L steps, each one of size $\epsilon$ and iterate between an half step for the momentum and a full-step update for the position.

$$\begin{aligned} \boldsymbol{\phi}_{t+\frac{\epsilon}{2}} &= \boldsymbol{\phi}_t - \frac{\epsilon}{2}\frac{\partial V}{\partial \boldsymbol{\theta}_t} && (35) \\ \boldsymbol{\theta}_{t+\epsilon} &= \boldsymbol{\theta}_t - \epsilon M^{-1}\boldsymbol{\phi}_{t+\frac{\epsilon}{2}} && (36) \\ \boldsymbol{\phi}_{t+\epsilon} &= \boldsymbol{\phi}_{t+\frac{\epsilon}{2}} - \frac{\epsilon}{2}\frac{\partial V}{\partial \boldsymbol{\theta}_{t+\epsilon}}. && (37) \end{aligned}$$

The leapfrog discretisation introduces small numerical error in the total energy calculation. The correction take the form of a Metropolis-Hastings step, in which the probability of accepting a proposal $(\boldsymbol{\phi}^*, \boldsymbol{\theta}^*)$ generated from $(\boldsymbol{\phi}, \boldsymbol{\theta})$ is $\min(1, \exp(H(\boldsymbol{\phi}, \boldsymbol{\theta}) - H(\boldsymbol{\phi}^*, \boldsymbol{\theta}^*)))$. In case of rejection, the previous values are used to initialise the new iterations. In practice, when using HMC two main parameters need to be tuned. Firstly, one need to chose the appropriate step size. Taking a look to the acceptance rate, it is possible to reduce or increase the value of $\epsilon$. Smaller step are more computationally expensive but precision

may benefit. On the other hand too small $\epsilon$ make it difficult to efficiently explore the target distribution. Instead, the best way to determine the appropriate length L of the simulation. one can take a look to the parameters auto-covariance function, increasing L to achieve more independent samples. Too long trajectories can erode computational effort, as the simulation exercise may generate loop, making the destination point the same as the beginning one. Once determined reasonable values for $\epsilon$ and L, desired sample from the target distribution can be obtained.
Generally, for MOPG, the parameter of interest are the one composing the corregional matrix B which can describe relationship among the outcomes. Let us call these parameters $\boldsymbol{\eta}$. In Hierarchical Bayesian models, one can set up different levels of latent variables. The conditional posterior of $\boldsymbol{\eta}$ can be computed as

$$p(\boldsymbol{\eta}|\boldsymbol{\theta}, \mathbf{y}, X) = \frac{p(\mathbf{y}|X, \boldsymbol{\eta}, \boldsymbol{\theta})p(\boldsymbol{\eta}|\boldsymbol{\theta})}{p(\mathbf{y}|X, \boldsymbol{\theta})}. \tag{38}$$

By maximizing the the denominator in (38) it is possible to obtain the maximum likelihood type II estimates $\boldsymbol{\theta}^*$ ,

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\mathbf{y}|X, \boldsymbol{\theta}). \tag{39}$$

Then, the marginal posterior of $\boldsymbol{\eta}$ can be approximated by conditioning on the estimates obtained by ML-II optimization

$$p(\boldsymbol{\eta}|X, \mathbf{y}) = \int p(\boldsymbol{\eta}, \boldsymbol{\theta}|X, \mathbf{y}) d\boldsymbol{\theta} \approx p(\boldsymbol{\eta}|X, \mathbf{y}, \boldsymbol{\theta}^*) \tag{40}$$

In this way one can focus the attention and computational burden only on the parameter of interest.
The main toolkit for the analysis and optimization is GPy, a Gaussian Process (GP) framework written in Python.[6]

# 5 Empirical Analysis

## 5.1 Covid-19 Vaccination program

Accumulating evidence suggests that vaccination against Covid-19 reduces the risk of severe complications (including death) and slows down the growth of infections. A relevant issue, however, is how much of the observed slow down in the spread and how many saved lives are attributable to fast and effective inoculation policies as opposed to more

[6] https://sheffieldml.github.io/GPy/

conservative programs. The United Kingdom has delivered one of the world's fastest vaccination campaigns, giving the first shot to about 67% of the adult population and a second to 50% by the end of June 2021, helping to reduce deaths and infection rates.

## 5.2 Data

Our data consists of weekly data points for different countries from $1^{st}$ March 2020 to $30^{th}$ June 2021. The date of the intervention $t_0$ is set up on the $31^{th}$ January 2021, since that is the first week in which the number of people that had the second dose surpassed 500,000. As mentioned above, the treated country for this study is the United Kingdom as its policy differentiated from the other European countries.

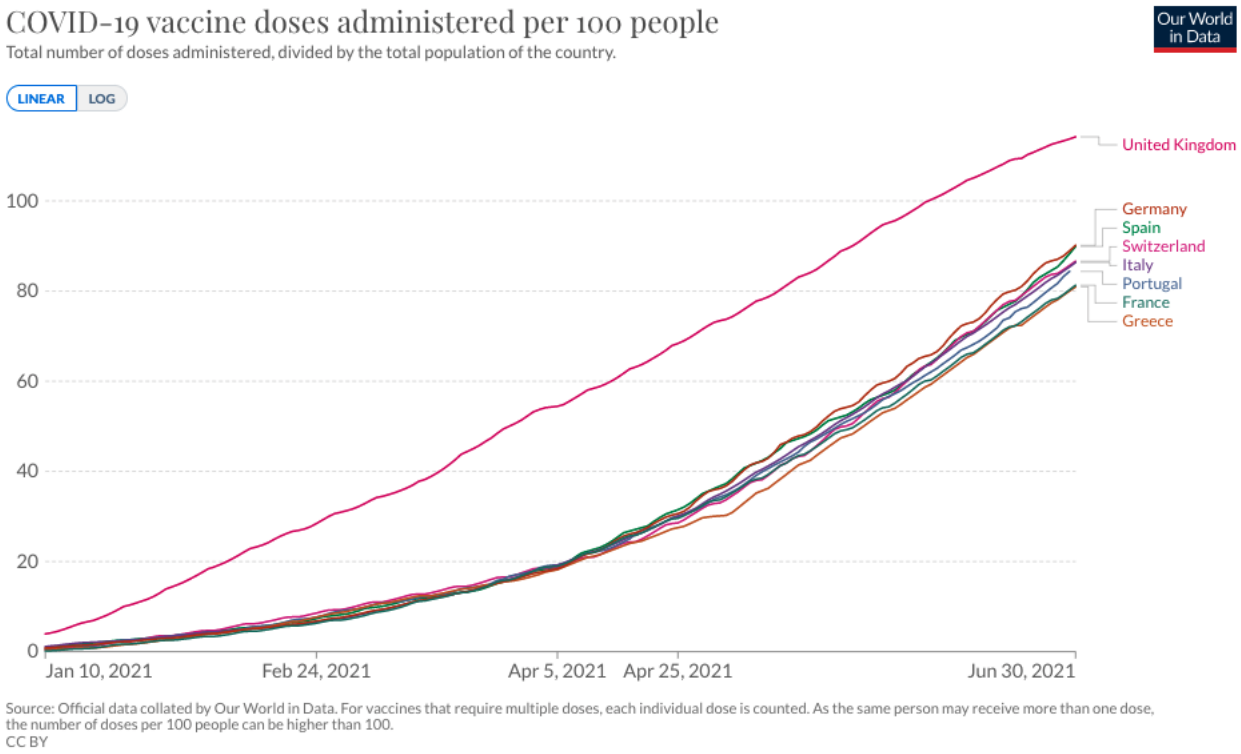



Figure 1: Covid-19 Vaccine doses administered per 100 people

We are going to focus our analysis on two different outcome variables. The first one is the confirmed Covid-19 deaths per million people. The variable is divided by the number of population for each country to obtain a continuous variable and then it is converted

on a log scale to better handle the asymmetry of the data arising from the absence of negative values. The second variable of interest is the estimate of the reproduction rate (R) of Covid-19. Generally, it measures the level of contagiousness of the virus and equals the expected number of cases directly generated by one case in a population where all individuals are susceptible to infection.
To respect treatment independence assumption, all covariates are considered unrelated to the intervention. They are: i) time trend, ii) Google Mobility Data and, iii) weekly number of Covid-19 tests. The first one is a variable that express time domain with $t = 1$ being the first time we dispose of an observation. Since we are working with heterotopic data, each of the outcomes variable is associated with different covariates with potential different start time.[7] This variable express a time trend, although it is not necessarily linear thanks to the flexible structure of the kernel function on the input space. *Google Mobility Report* is a publicly available dataset that express how visits at different places changed overtime, compared to a baseline. The venues covered by the dataset are: grocery and pharmacy, parks, transit station, retail recreation, residential and workplaces. To reduce the dimension of the input space we perform Principal Component Analysis (PCA) on each country and we use the first principal component as a single variable which represent country mobility. On average more that 80% of the variability of the dataset is explained by this factor, making pointless the use of other components. The last variable expresses the weekly average Covid-19 tests per 1000 people.

## 5.3 Results

### 5.3.1 Weekly death per million people

We start by applying the DTW algorithm to select the most similar (lower "distance") countries with respect to the UK. This process restrict the pool of European candidates to the following countries: 'Italy', 'Netherlands', 'Ireland', 'Switzerland', 'Germany', 'Portugal', 'Poland', 'Greece', 'Croatia', 'Belgium'. Given these countries we first focus on forecasting performance before the intervention $t_0$. We fit different model as outlined in the methodologies and obtain that the best results are achieved using five countries, namely 'Italy', 'Netherlands', 'Ireland', 'Portugal'.

[7] For example in the *number of deaths* exercise we have that Italy (the country whose reports are the earliest) first reporting date is '2020-03-15', thus the time value will be assigned as 1. United Kindom first data report is instead '2020-04-19' and it will be assigned a value $t = 5$, as this date is 5 weeks forward.

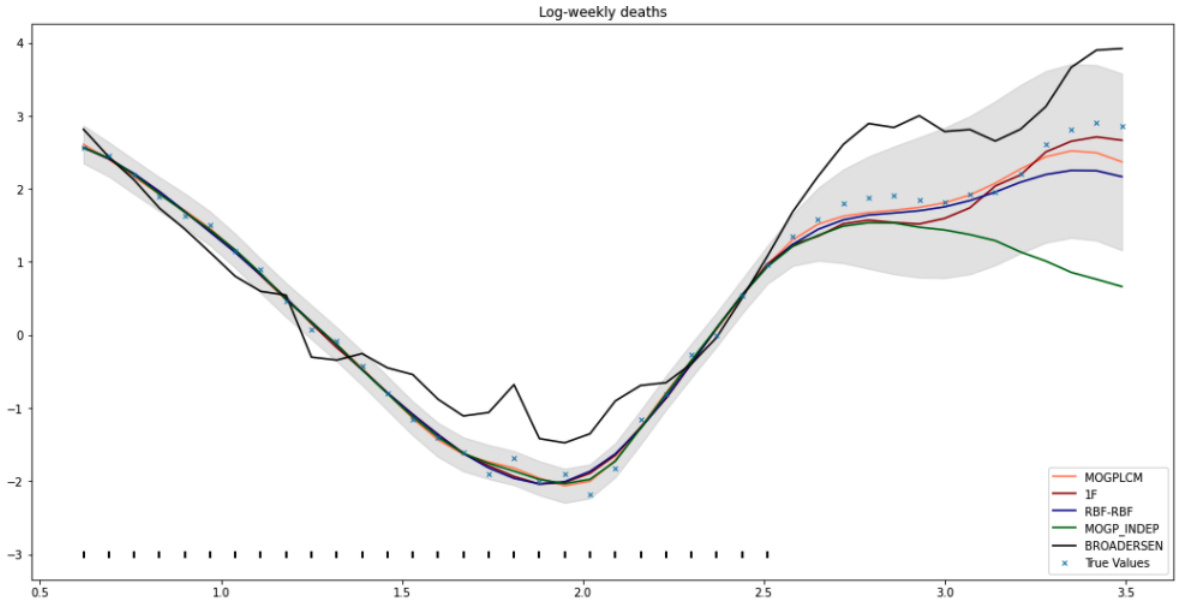



Figure 2: Predicted log weekly deaths per 1000 people for the different models. The blue cross represent observed data, and each line correspond to a specific model. The black tickers at the end of the graphs show which data points have been used to train the model. Gray shaded area represent 95% prediction confidence interval of the SLFGP

| | SLFGP | 1FGP | 2RBFGP | INGP | BCI |
|---|---|---|---|---|---|
| MSE | **0.8189** | 0.8274 | 1.2754 | 4.2811 | 3.1693 |
| logS | **0.2047** | 0.4252 | 0.3115 | 1.1334 | 5.5936 |
| ES | **0.6704** | 0.7641 | 0.8827 | 2.7793 | 2.6316 |

The table indicates that the Semi Latent Factor Gaussian Process model is the best one, having a smaller value on each of the measures.The other two models (1FGP and 2RBFGP) perform well overall but slightly worse compared to the base model. The INGP is the worse GP model since it cannot rely on the part of data after-intervention coming from the other countries, thus converging so the average number of weekly deaths. BCI also does not achieve the same result as its competitors.
Once the model is established, it is possible to fit the model to the whole data set. Now, UK dataset $\{(y_{i,t}, \mathbf{x}_{i,t})\}_{t_i}^{t_0}$ will be restricted to all points before intervention $t_0$ while for the other countries $j \neq i$ the whole dataset $\{(y_{i,t}, \mathbf{x}_{i,t})\}_{t_j}^{T_j}$ is used. Using *type II* Maximum likelihood as in (28) one can obtain the optimal hyperparameters for the kernels. The

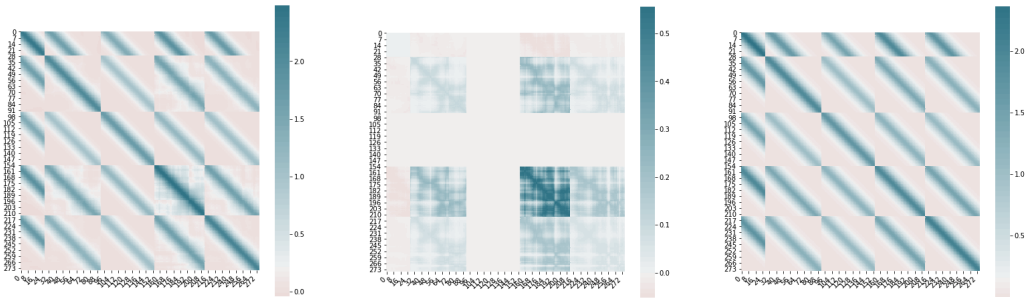


Figure 3: The figure on the left is the total variance $\mathcal{K}$ as in (17), while the other two represent the component Kernel on the covariate space $\mathrm{B}_1 \otimes K_{rbf}(\overline{X}, \overline{X})$ and time space $\mathrm{B}_2 \otimes K_{Mat}(t, t)$, respectively

estimated kernel matrices are of dimension $mT \times mT$ where $T = \sum_{i=1}^{m} T_i$ and represent the variances and covariances of each data point. The kernel on the left is $\mathcal{K}$, which is the sum of the outer products of each kernel times the coregionalization matrix B. Each of the $m$ blocks represent the variance-covariance matrix of each process $f_j(\mathbf{x}_j)$. The upper left block is the one corresponding to the UK and it is smaller compared to the others, as fewer data points are employed. A first visual inspection of the $\mathcal{K}$, one can note that the time component accounts for the majority of the variability as values spans from 0 to 2. Furthermore, we see that mobility data and number of test were very important variables to explain Ireland (fourth country) but almost no influential in the Netherlands (third country). However, we see how the independent variables of other countries explain very little for the UK (top row/column of the matrix). The time-domain coregionalization matrix is more homogeneous and we can see that UK weekly death followed a pattern more similar to Ireland (slightly higher values of $B_{2,[4,4]}$) compared to other countries. As the kernel is stationary, all that matters is the relative distance in time between countries. Lockdown measures, vaccination programs, etc do not have to match, and the relationship between the main variable and lagged/forward version of the control units are still captured by the model, even if they are not linear.

Now, let us focus on the estimate of the causal effect. As shown in 3, the GP model provided a great fit for the pre-intervention period. Following the beginning of the vaccination campaign, observations start to diverge from the counterfactual predictions: the actual number of weekly deaths, represented by the blue crosses, was consistently lower than what would have been achieved at slower vaccination rates. Subtracting observed data from predicted, as is shown on the right part of 4 produces the posterior estimate of the differential effect achieved by the campaign. The top right graph gives an idea of the cumulative log number of deaths for the UK compared to control countries. Before the

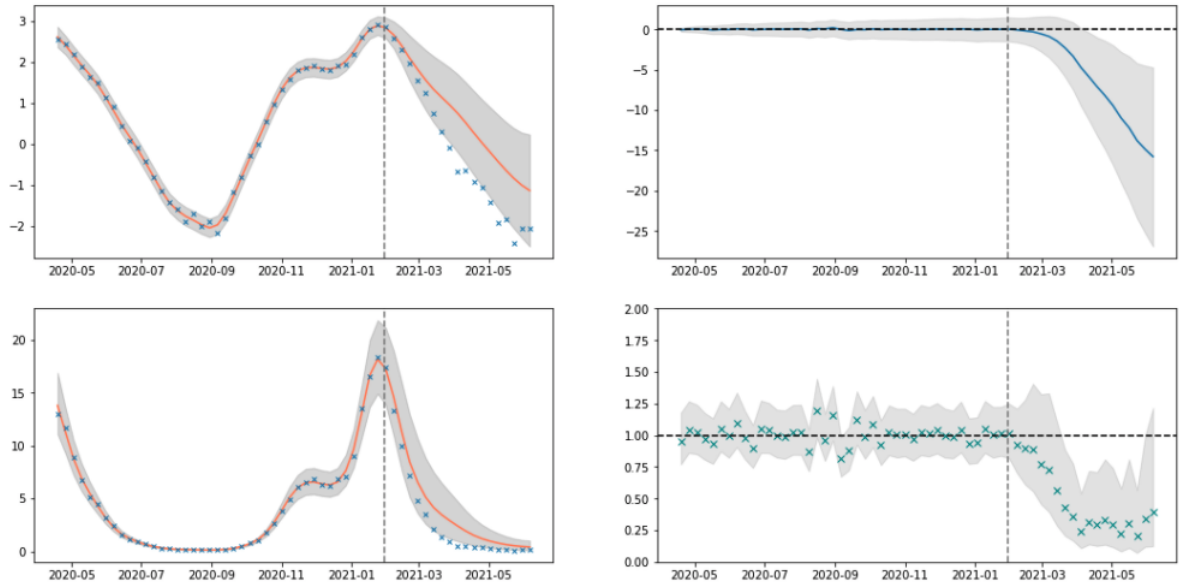



Figure 4: The graphs on the left indicate the log weekly deaths on top and the level on the bottom. Data provided to train the model is at the left of $t_0$, the gray vertical line. Orange line represent model predicted average of (log) weekly death with 95% credible intervals in gray. True values are in blue. On the right part, the top graphs show the cumulative effect of log weekly deaths $\mathcal{T}$while the bottom display pointwise multiplicative causal effect $\tau_t^*$.

intervention the cumulative difference was statistically non-significant, .i.e the differential number of deaths was the same among the countries. After $t_0$ however, the cumulative effect starts going down, reaching a value significantly lower than 0. In the UK the counter-factual (log) number of deaths was higher than the actual values, denoting how the vaccination campaign managed to save lives. The bottom right graph shows instead the pointwise estimate of $\tau^*$, the multiplicative causal effect with 95% credible regions. In the period spanning from April and May 2020, the average relative effect was 28.32% [11.24%, 71.8%], that is for every 100 deaths per million people in non-treated countries were corresponding around 28 in the UK. To understand the average effect right after the campaign one can calculate the average effect as in (5). Over the whole period, on average, of two people dead by Covid-19 in the rest of Europe was corresponding just one death in the UK (that is, the ratio is 52.83% [30.19%, 83.14%])

### 5.3.2 Weekly infection rate $\rho$

The same analysis is run for the weekly infection rate, to measure if a different vaccination campaign is actually producing slower rates of infectiousness. [8]After Applying DTW algorithm to UK $\rho$ we are left with 10 countries, namely 'Portugal', 'Spain', 'Belarus', 'Sweden', 'Ireland', 'Russia', 'Norway', 'Luxembourg', 'France', 'Denmark'. Then as we did for the weekly deaths, we focus on the period before the intervention to find out which combination of five countries produces the best prediction accuracy. Half of the selected countries selected ('Portugal and "Ireland') are in common with the previous analysis while 'France' and 'Denmark' take the spot as new control units.

| | SLFGP | 1FGP | 2RBFGP | INGP | BCI |
|---|---|---|---|---|---|
| MSE | **0.3597** | 0.4679 | 0.4027 | 2.1434 | 0.5273 |
| logS | **-0.6488** | -0.6303 | -0.7744 | 0.8579 | -0.6339 |
| ES | **0.2897** | 0.3340 | 0.2903 | 1.4377 | 0.3623 |

As the table shows, lower values of each of the three metrics are achieved for the main model SLFGP. Once again, independent GPs (INGP) are not able to capture post-intervention information from the control variables, making the prediction converge to the mean. BCI, also performs worse compared to its non-linear counterparts, even in the train section of the data.
We can now employ *type II Maximum likelihood* on the whole data set to find the optimal hyper-parameters of the kernel. Given that the preproduction rate is an estimate itself, it can be affected by many sources of variability induced by the data or the model used. To take into account this effect, we bound the likelihood variance (observation error) to a minimum value of around 0.01, which corresponds to 5% of countries' reproduction rate variance over the observed period. Differently for the weekly number of deaths, the time component does not seem to play a prevailing role in defining $\rho$ dynamics. On the other hand, google mobility data, i.e. which places people were visiting during the period was effective in predicting how the contagiousness would change. As the variable is a principal component transformation, little interpretation can be given to single venues. With regards to the time component, one can note that there is mainly a contemporaneous effect, as the lengthscale $\ell$ of the Matern kernel (12) is lower compared to the one estimated in 5.3.1. Decreasing the length parameter reduces the banding, as points further away from each other become less correlated. This means that data points have zero covariance with the lagged version of both the treated variable and control units. This

[8]We removed the number of test variables as it did not affect results, making the optimization more challenging.

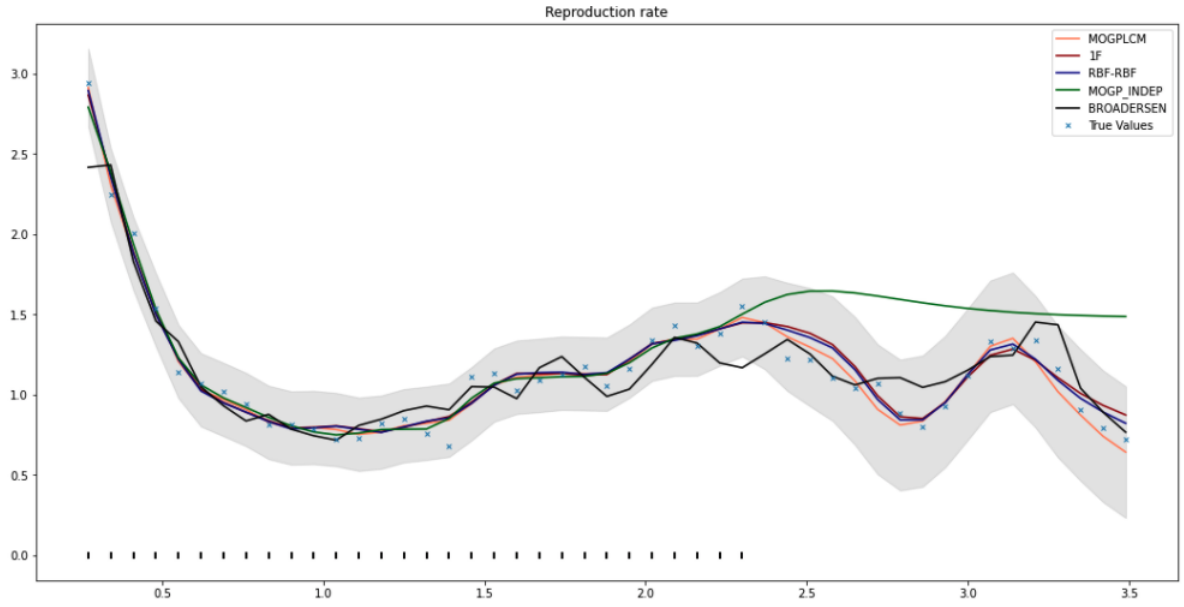



Figure 5: Predicted rate of infectiousness for the different models. The blue crosses represent observed data, and each line correspond to a specific model. The black tickers at the end of the graphs show which data points have been used to train the model. Gray shaded area represent 95% prediction confidence interval of the SLFGP

effect is even minimized for 'France' (fifth country) as the estimate of $B_{2,[5,5]}$ is about half the one of other countries. The causal effect is then estimated on the available dataset, restricting the UK on data before $t_0$. As shown in 7, before the vaccination campaign the model provide a good fit of the data. Afterward, counterfactual predictions (orange lines) initially follow the main trend of observed data but then they start to deviate marginally. Yet, the variability surrounding the estimate is too broad to confirm any causal effect in the data. It's interesting to note that in the very final period of the analysis, starting May 2021, the model predicts a stable reproduction rate. However, the data seems to diverge positively to higher values. The main reason for this discrepancy is mainly due to the spread of the *delta* variant. This more contagious version of Covid-19 represented 73% of UK cases by the end of May 2021. Nevertheless, on the period following the vaccination campaign the average additive causal effect $\overline{\tau}_i$ is estimated and equals 0.0074 with 95% credible interval [0.1621, -0.15], thus no reduction is detected. [9]

[9]On the contrary, it is slightly positive, although not statistically significant

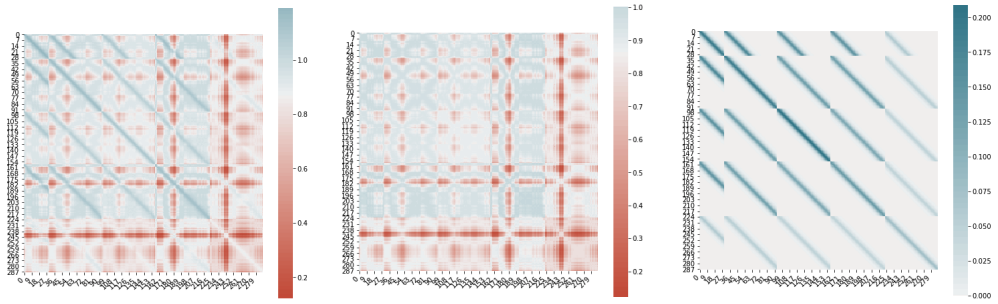


Figure 6: The figure on the left is the total variance $\mathcal{K}$ as in (17), while the other two represent the component Kernel on the covariate space $\mathrm{B}_1 \otimes K_{rbf}(\overline{X}, \overline{X})$ and time space $\mathrm{B}_2 \otimes K_{Mat}(t, t)$, respectively

# 6 Conclusion

A freshly growing literature on applied causal inference indicates an increasing interest in evaluating the incremental impact of market interventions and policies. With this paper, we propose a novel approach to obtain the counterfactual prediction of the unobserved market outcome. We employ a Bayesian Machine Learning technique, based on Gaussian Processes, whose main features are discussed below.

The prevailing literature of dynamical Bayesian causal model revolves around Brodersen et al. (2014), whose approach is based on state-space models, which promptly lend themselves to posterior inference. However, since closed-form solution to the posterior is challenging to obtain, the authors resort to stochastic approximation, using MCMC.

In the general form of Gaussian Processes, causal effect posterior evaluation can be instead derived analytically. When this is not possible - for example when using transformation of variables or cumulative measures - one can employ GP sampling procedure. As a GP is fully characterized by its mean (generally 0) and its variance (the kernel), the process is straightforward.

Furthermore, state-space models impose some restrictions on the dynamic evolution of the states, first of all, linearity. With GPs, the input $\mathbf{x}$ is transformed into a feature vector $f(\mathbf{x})$ through some non-linear mapping, dictated by the structure of the kernel. On a time series, the degree of correlation between a variable and its lag is given by the relative time distance. When using a Matèrn kernel, it can be shown that the covariance matrix of the kernel gives rise to a particular form of a continuous-time AR(p) Gaussian process (Rasmussen and Williams, 2006). At the same time, nothing prevents it from using a more complex structure - such as periodic, linear, etc., or a combination thereof - to better fit the time curve. Furthermore, the linearity assumption with exogenous

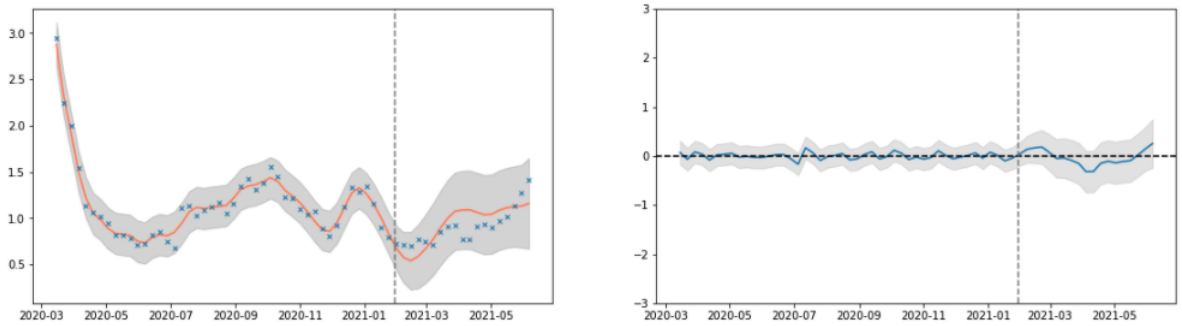



Figure 7: The graphs on the left indicates the evolution of the reproduction rate in UK. Data provided to train the model is at the left of $t_0$, the gray vertical line. Orange line represent model predicted average of $\rho$ with 95% credible intervals in gray. True values are in blue. On the right part, the graph display pointwise additive causal effect $\tau_t^*$, with relative 95% bands in gray.

regressors embodied in state space can be relaxed by adopting an appropriate kernel architecture. Machine Learning tools, such as cross-validation can help decide which one better describes the data.

Another important improvement that GPs put forward is an *heterotopic* configuration of data, i.e. each output has different training set with a potentially different number of samples. This approach plays a crucial role in causal analysis since generally one has to discard all information after the intervention period to train the data, generating a non-negligible loss of data. In addition, no time matching is needed as the model understand the relationship among the potential outcome and the explanatory variables for each unit, independently if these variable match in absolute time.

Lastly using GPs one can easily quantify uncertainty around a measurement or prediction, since every data point possesses a defined distribution. This promotes direct estimation of the causal effect distribution, means, and quantiles.

To set this model in practice, we dealt with measuring the efficacy of vaccination policies throughout Europe. In particular, we analysed how the United Kingdom faster inoculation campaign affected the cumulative number of deaths and rate of contagiousness measured by the reproduction rate, all things being equal. The study suggested that, indeed, vaccination help prevent deaths, since on average, in the first semester of 2021, every death in the UK related to Covid-19 was corresponding to two deaths in the rest of Europe. Different results were obtained for the reproduction rate as no statistically significant evidence was found to justify that vaccines reduce the number of cases directly caused by an infected individual. However, this has to be considered in light of the new

and more infectious variants that started spreading over the continent at the end of our sample period.

# References


Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(1):113–132.

Aglietti, V., Damoulas, T., Álvarez, M., and González, J. (2020). Multi-task causal learning with gaussian processes. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6293–6304. Curran Associates, Inc.

Alaa, A. M. and van der Schaar, M. (2017). Bayesian inference of individualized treatment effects using multi-task gaussian processes. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alvarez, M. A., Rosasco, L., and Lawrence, N. D. (2012). Kernels for vector-valued functions: a review.

Bonilla, E. V., Chai, K., and Williams, C. (2008). Multi-task gaussian process prediction. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2014). Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*, 9:247–274.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in r: The dtw package. *Journal of Statistical Software*, 31(7):1–24.

Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*, 90:1–37.

Liu, H., Cai, J., and Ong, Y.-S. (2018). Remarks on multi-output gaussian process regression. *Knowledge-Based Systems*, 144:102–121.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096.

Menchetti, F. and Bojinov, I. (2020). Estimating the effectiveness of permanent price reductions for competing products using multivariate bayesian structural time series models. *Harvard Business School Working Paper*, 21-048.

Murphy, K. P. (2013). *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.].

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, New York, NY, USA, second edition.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press.

Sävje, F., Aronow, P. M., and Hudgens, M. G. (2019). Average treatment effects in the presence of unknown interference.