Kassandra's predictor

Pantazis Y. and Vernardos G.

February 17, 2021

Abstract

We build and then train an ensemble of predictive models for each country/region. All models are regression models for the rate of change of new cases at the logarithmic scale. The rate of change is modeled as a function of some of the interventions resulting in a sparse regression problem. Each element (model) in the ensemble has the same functional form but it is computed with different estimation methods, as well as on different time horizons. Our final predictions are computed as the median of all models' predictions. An additional advantage of having multiple predictive models is their ability to estimate confidence intervals and quantiles.

1 Model Description

We define an archetypal predictive model as follows. Let $c(t)$ be the time-varying variable representing the new cases for a country/region, and $z(t) = \log\big(c(t)\big)$ its logarithm. The use of the logarithm essentially implies that we assume multiplicative contributions on the number of new cases. Since $z(t)$ is varying significantly due to the reporting policies of each country/region, we smooth it with a moving average filter with length L_f. When $L_f = 7$ we have a 7-days moving average but larger values have been also applied to further smooth the time-series. Our target variable is the time difference of $z_s(t)$ which is the smoothed version of $z(t)$.

The time-varying predictor variables are characteristic functions of the intervention plans plus the intercept, i.e., a constant value which models the infection rate. We do not include any additional factor. The characteristic functions are also smoothed with a 14 days filter which anticipates the gradual and slightly-delayed impact of an intervention. The delay is also a result of the virus incubation time. We denote the smoothed intervention plans by $\phi_l(t)$ with $l = 1, ..., L$ and the constant term with $\phi_0(t)$.

Figure 1 shows the target variable (lower left panel) and some of the predictor variables as a function of time for Canada.

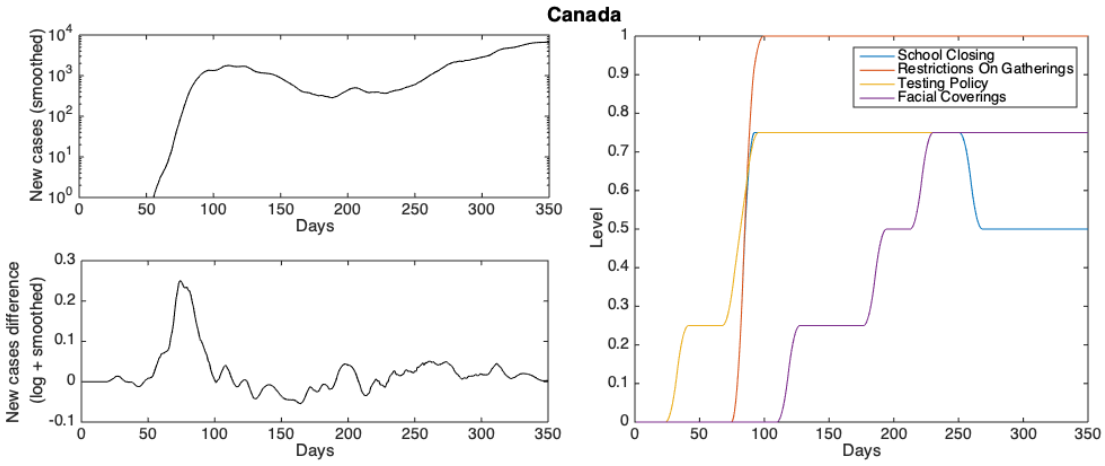

Figure 1: Canada's new cases (left) and indicative intervention plans (right). We essentially model the time-series in the left bottom panel with a weighted sum of the interventions in the right panel.

We deploy a linear predictive model with unknown coefficients, a_l with $l = 0, ..., L$ given by

$$z_s(t) - z_s(t-1) = \sum_{l=0}^{L} a_l \phi_l(t) \ . \tag{1}$$

The use of a linear model is motivated by robustness concerns and the fact that extrapolation, i.e., future forecasts, typically diverges significantly with a non-linear, complex model. Another reason for using such simple linear models is that the data are not adequate and in many cases prone to errors.

2 Model Training

We use two different estimation approaches in order to compute the coefficients of the predictive model shown in equation (1). One estimation approach is Orthogonal Matching Pursuit (OMP)[1] which produces a sparse solution meaning that some of the coefficients will be exactly zero. Figure 2 shows the time-series obtained from the OMP algorithm on Canada's reported new cases (red lines) and compared with the ground truth (black lines).

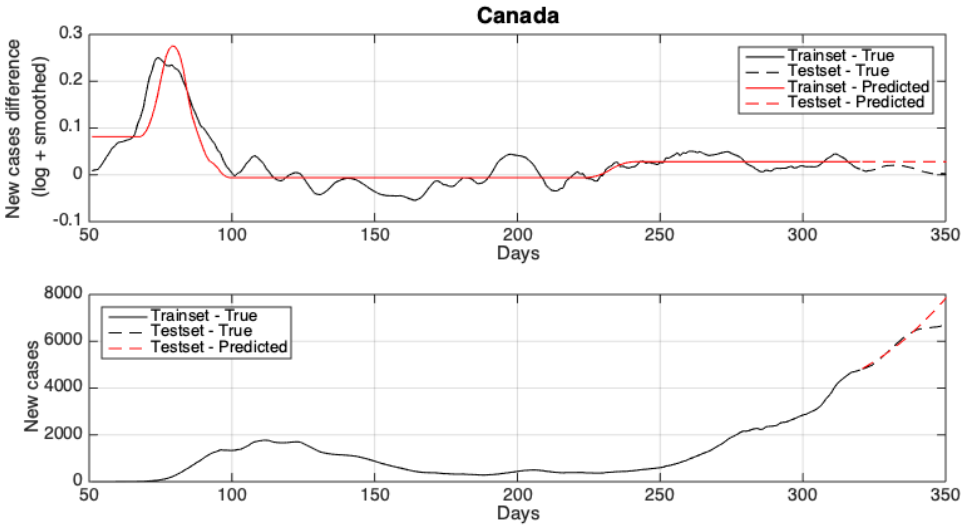

Figure 2: Canada's regression outcome when OMP is used as an estimation method (upper panel) and the corresponding predictions for the future new cases (lower panel).

The other estimation approach is an exhaustive search of candidate models using a validation set to determine which models perform best on future periods. For the exhaustive search approach, we estimate all combinations of models with $1, 2, ..., L$ interventions in it using Regularized Least Squares (RLS) [2]. Essentially, we estimated hundreds of models for each country/region and screen out the ones with the best performance. Indeed, the ranking of the models is performed using a validation set where the mean absolute error is computed. We keep the 50 best performing models and from those 50 best-performing models we keep both the mean and the median predictions.

Overall, we created 37 models for each region/country and we report the median value of them. We train those models by varying the training set horizon as well as the smoothness applied on the new cases. First, we observe that the results were significantly different depending on the time horizon used for training the models. Therefore, we train models for three different horizons of time: one that spans all the covid-19 pandemic from mid February and onward, one with starting date in early June and one with starting date at end of August. Secondly, we observe that smoother time-series where fitted with greater accuracy. Therefore, apart from a 7-days moving average, we apply a 14- and 21-days moving average as well. For some horizons, we also apply 28- and 42-days moving averaging of the time-series.

[1] https://en.wikipedia.org/wiki/Matching_pursuit

[2] https://en.wikipedia.org/wiki/Regularized_least_squares

Finally, an important feature of the training process is the imposed restrictions on the coefficients. For instance, the infection rate cannot be negative while restrictions on gathering or school closing cannot have a positive impact on the number of new cases. This sanity check is necessary due to the discrepancies and inconsistencies of the available reported data. All prohibitions as well as facial coverage are considered to have a negative impact to the number of new cases. However, the role of testing policies and contact tracing is less clear because it initially causes a positive increase of new cases before it eventually assists with the containment of the virus. We made the difficult decision to impose a positive impact on the number of cases from these two intervention policies, therefore, the corresponding coefficients are not allowed to be negative.

3 Results and Discussion

An interesting outcome from the sparse regression analysis is how many times an intervention has been overall selected. Figure 3 gathers the occurrence frequency of each intervention. It seems that some interventions are selected more often for the description of the observed logarithmic new cases.

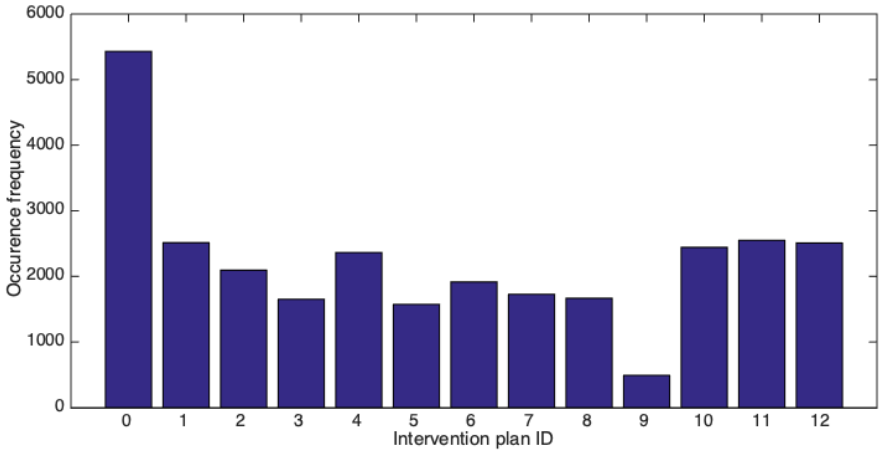

Figure 3: Occurrence frequency for each intervention. The five most frequent interventions that were picked up by the estimation algorithms were Contact Tracing (11), School Closing (1), Facial Coverings (12), Testing Policy (10) and Restrictions On Gatherings (4). The intervention that was picked the least was Public Information Campaigns (9). Note that 0 corresponds to the constant positive rate of infections.

We observe that there are many uncertainties in the data which make even the modeling of previous new cases very challenging. Some typical examples include countries/regions in the dataset with different reporting policies, special political agendas and different enforcement rules for the intervention plans. Additionally, there are mismatches/errors in the dataset itself since strictness levels might be misunderstood due to cultural differences and the lack of a global consensus.

Our focus is primarily concentrated to find which interventions can explain better the results so far. Our thesis is that knowing which interventions are effective and how much they affect the development of the spread, we could forecast more reliably the long-term horizon of future new cases as a byproduct. Therefore, we have slightly sacrificed the short-term accuracy for a more robust, reliable and explainable model. No other data source is used even though it is expected to improve the results.

Finally, we would like to highlight that the computational time for the training of all models for all countries/regions on a standard laptop with an i5 CPU is 30 minutes. In total, we train more than 6 million models in order to produce the ensemble forecasts that we provide. This implies that the training of a single model is on average less than 1 millisecond.