DS-GA 1013 Mini Project

Jing Tao
jt3149@nyu.edu

Jianfei Xue
jx898@nyu.edu

Daiheng Zhang
dz2266@nyu.edu

March 10 2022

1 Introduction

The outbreak of covid-19 has already been 2 years since 2020. An important task is to evaluate the real-time infectivity of the virus, which is crucial for policy making and vaccination development. In this project, we used the COVID-19 Statistics dataset from *Our Word in Data*[1], which includes 67 different COVID-19 daily statistics of countries in North America from March 2020 until now. It includes pandemic statistics like new cases, new death, vaccination and geographical statistics like GDP and HDI. Hence we are interested in understanding the pandemic from a data-driven way by doing singular value decomposition and inducing the potential factors and fitting the data to regression models to analyze the underlying factors, current situations and future trend. In particular, we will use data of United States for all the experiments.

2 Data Exploration

In the dataset, there are around thirty features on pandemic but most of them are derived from original data for example "new cases per million" is just "new cases" divided by population and "total cases" is just adding up "new cases" till date. Therefore, we choose only the original, unmodified observations: new cases, new deaths, reproduction rate, daily ICU patients, daily hospital patients, and positive rate of testing. And since each of the observations reveal some information about the pandemic, it makes sense to center the data by column mean and do a Singular-Value-Decomposition and see how the singular vectors look and what they imply.

date	new_cases	new_deaths	reproduction_rate	icu_patients	hosp_patients	positive_rate
2020-03-05	77.0	1.0	3.62	0.0	0.0	0.000
2020-03-06	53.0	2.0	3.56	0.0	0.0	0.000
2020-03-07	166.0	3.0	3.58	0.0	0.0	0.108
2020-03-08	116.0	4.0	3.46	0.0	0.0	0.112
2020-03-09	75.0	1.0	3.32	0.0	0.0	0.110
...	...	...	...	...	...	...
2022-02-28	96853.0	2096.0	0.62	7551.0	37860.0	0.039
2022-03-01	47065.0	1694.0	0.60	7326.0	36193.0	0.036
2022-03-02	53725.0	2098.0	0.60	6982.0	34703.0	0.035
2022-03-03	49232.0	1745.0	0.61	6728.0	32935.0	0.033
2022-03-04	51565.0	1691.0	0.61	6224.0	30692.0	0.033

Figure 1 is the United States pandemic data shown in tabular format, each column is a time series of how one observation changes over the span of two years. Each row is the observations at a given date. Therefore, the resulting left singular vectors will show factors that affects the observations over time and the signs of entries of the resulting right singular vector tell us each corresponding feature reacts to this factor positively or negatively.

Figure 1: USA pandemic data in tabular format

As illustrated in Figure 2, this is the graph for the first left singular vector plotted against the date. Our interpretation is that this factor is the "intensity of pandemic" since all the entries in the

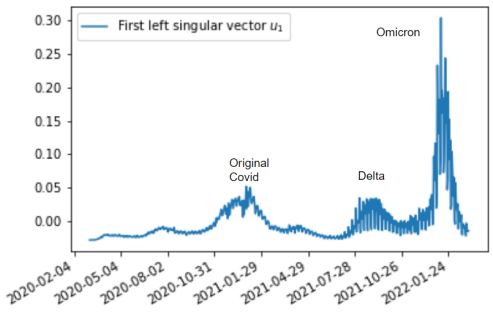

Figure 2: first left singular vector against the date

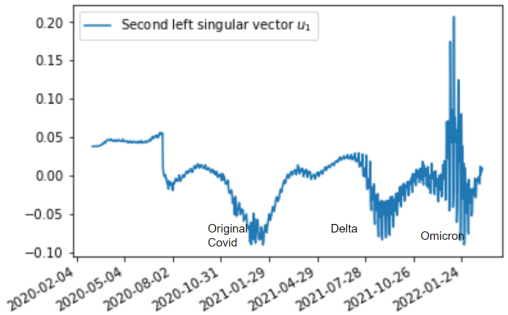

Figure 3: second left singular vector against the date

right singular vectors are positive for the features. That implies that when the intensity is high all the observations go up. Logically, that makes sense. Also if we look at those local maximums and their corresponding dates, the first maximum is roughly when the original covid outbreak occurred, and the two following correspond to the Delta and Omicron outbreak. These information matches with our guesses.

The graph shown in Figure 3 is the second left singular vector plotted against the date. Here we see a different pattern and also entries in the right singular vector now have different signs. The entries for new deaths, ICU patients and hospital patients have negative signs. In essence, those three features convey how deadly the virus was. Again we include the date information, the local minimum that happened around the end of year 2020 can be identified as the original covid outbreak. The negative value in the left singular vector as we observed in the graph multiplies the negative entries of the three features in the right singular vector and scale the value with the singular value, we get a huge rise on the hospital and ICU patients observed and daily deaths reported. The result implies that the original covid was most deadly to people. We can reason that people were not vaccinated back then and had no pandemic experience before. Similarly we get that Delta variant was less deadly and Omicron was the least deadly among the three. Those findings match with our real life understanding about these viruses, by the time Delta arrived, most people in United States were vaccinated and the vaccination was very effective against Delta. Omicron just in its nature does not cause severe symptoms most of the time and was least deadly compared to other variants.

3 Predicting the $\mathcal{R}$ number

The reproduction number $\mathcal{R}$, which is defined as the average number of secondary cases produced by a primary case, is a critical metric measuring current infectivity of a virus. Analytically, it is computed by the SIR ODE model[2]. However, in practice, it is hard to compute the true reproduction number due to the lack of secondary cases data since it is difficult to track infections. Hence we are interested in estimating the reproduction number using some accessible statistics. Here, we applied regressions to the pandemic data of the United States to predict the daily reproduction number given today's observed pandemic features. As mentioned before, the pandemic data contains over 30 features, so we need to do some feature engineering. We first drop the reproduction number feature since it is our label. We noted that there are also missing values in the feature data, and most of them happened in the first several rows, which makes sense since there is no available pandemic measures in the very beginning of the outbreak. So we just simply drop

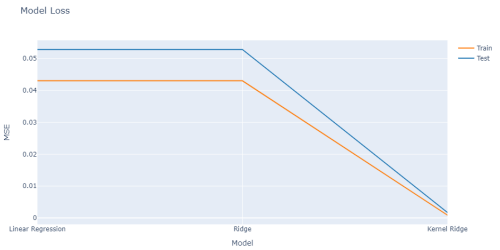

Figure 4: Loss

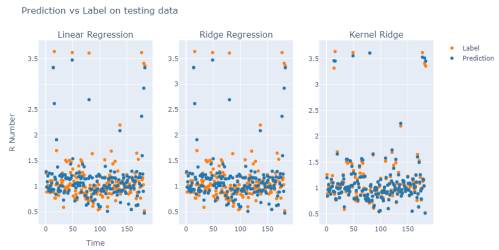

Figure 5: Predictions and Labels

those rows as they contribute no information. We observed that the magnitude of some features differs dramatically with others, so we normalized the data by subtracting mean and dividing by standard deviation. Then we fed the normalized data into Lasso regressor with cross validation (LassoCV) for feature selection. Note that Lasso regression leads to sparsity and it could learn the coefficient for irrelevant or redundant features to be 0. By this criteria, we selected 16 features corresponding to non-zero coefficients output by Lasso regressor, resulting in a $\mathbb{R}^{730\times 16}$ data matrix. After randomly splitting the data into training and testing set (75% train and 25% test), we fed the data to three different models we experimented: regular linear regression, ridge regression with cross validation (RidgeCV) and kernel ridge regression with polynomial kernel of degree 2. Figure 4 shows the mean squared error on both train test set and Figure 5 shows the testing results of the three models.

4 Analysis of Regression Results

Observing from Figure 4 and Figure 5, the kernel Ridge outperformed the other two in both losses and predictions, although Linear Regression and Ridge also have relatively low prediction error. There are two plausible interpretations: (1) The kernel Ridge can potentially overfit. Due to the random split of train test data, the kernel Ridge may "memorize" the nearby points of the points we are predicting, and for time series, nearby points may share similar labels, which makes our kernel Ridge behaves like a K-Nearest Neighbors model. (2) Since the real $\mathcal{R}$ number is derived from ODE, then it makes sense to use polynomial feature mapping, which captures hidden information about the $\mathcal{R}$ number. Hence boosting the accuracy of the kernel Ridge.

Besides kernel Ridge, however, the regular Linear Regression and Ridge Regression also performs relatively well. The regularizing parameter for Ridge is $\lambda = 10^{-5}$ from cross validation. This means the regularizing term did not penalize the model much and it behaves like the regular linear regression, which is supported by our Figures that the losses and predictions of linear regression and ridge regression are very similar.

5 Conclusion and Further Improvements

Overall, regular and Ridge Linear Regressor performs well in estimating the current-day reproduction number. In the next steps, instead of just the United States, we are interested in cross-country analysis by clustering countries with similar pandemic situations and predict the trend of infection from each other. Also another improvement is changing the train test split method and feature engineering to test if the model could predict the future reproduction number.

References

[1] Our World in Data. "Coronavirus Disease (COVID-19)—Statistics and Research". In: (2020). URL: https://ourworldindata.org/coronavirus.

[2] William Ogilvy Kermack and A. G. McKendrick. "A contribution to the mathematical theory of epidemics". In: 115 (1927). ISSN: 2053-9150. URL: https://doi.org/10.1098/rspa.1927.0118.