Predicting and Analyzing Growth and Prevalence of COVID-19 Cases in the United States

Anirudhan Badrinath

***Abstract*— As COVID-19 spreads rapidly through the United States, the study of which indicators most influence its growth and presence is quite important in "flattening the curve". As such, the goal of this computational project is to classify both useful and redundant geographical, social, and medical indicators of the prevalence and growth of COVID-19 in the United States as a function of person-to-person contamination in both urban and rural areas by making use of regression and regularization techniques, dimensionality reduction through principal component analysis, and exploratory data analysis. In discovering accurate and unique predictors with exploratory data analysis and PCA, we are able to construct a reliable linear regression model to make use of various types of indicators to predict the prevalence and growth of COVID-19. Throughout the analysis, it is determined that while many demographic factors such as population density or Medicare usage are tantamount to predicting the number of COVID-19 cases, there are just as many that are redundant such as pre-existing medical conditions and some age-related factors. It is determined that with just 3 days of vectorized data of the number of cases, we are able to reliably extract the entire evolution of COVID-19 in the United States using PCA, and moreover, with this information, we simply use around 20 features for linear regression to reliably predict the number of COVID-19 cases for any day with reasonably good accuracy ($R^2 \approx 0.72$, with low cross-validation error).**

I. Introduction

This project explores primarily the underlying relationships between different types of indicators that could be used to predict the growth and prevalence of COVID-19 through a variety of statistical techniques: dimensionality reduction, regression, and exploratory data analysis. In other words, the question explored throughout this project is: *what are the more useful as well as the more redundant geographical, sociopolitical, and medical indicators for the prediction of both the prevalence and growth of COVID-19 cases in the United States?* The data obtained contains geographical, sociopolitical, and medical information about individual counties, time series for COVID-19 cases and deaths, and information about states. Of those, we will use a combination of the county information along with case information to explain the growth and spread of COVID-19 in the United States.

We attempt to focus on two separate tasks: we classify which of the available data are useful and redundant in the prediction task, and we construct and improve a linear model that reliably predicts the number of COVID-19 cases on any given day. The first task is accomplished through the next two sections in which the data is initially explored manually and then in the following section using PCA. The second task is accomplished in the next section when a linear model is engineered and trained through feature engineering, regularization, and some randomized algorithms.

II. Exploratory Data Analysis

Before exploring the relationships between the variables through the lens of regression, the relationships between pertinent variables are explored manually after some data cleaning operations to merge the dataset about the counties with the time series describing the number of cases per day. In the former dataset, there exist many descriptions of geographical, social, medical and sociopolitical statistics about individual counties, which are merged together with the latter time series describing case information for each day starting in January for each county.

The clearest indicator based on intuition alone would be related to the population density given the transmission of COVID-19 through droplets: a higher population density should indicate a higher rate of growth and a greater prevalence of cases given even stricter regulation.

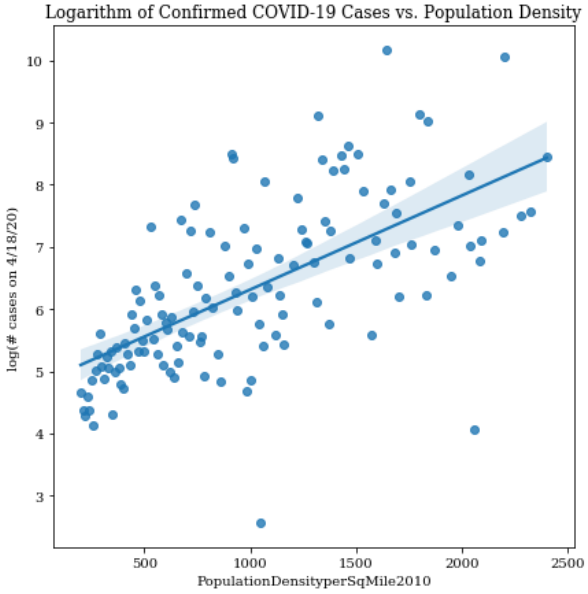

Fig. 1: The natural logarithm of the number of COVID-19 cases as of 04/18/20 are plotted against the population density per square mile as measured in 2010 for each county with 95% CIs, showing that the logarithm of the prevalence of COVID-19 monotonically increases linearly with population density. As the number of cases are exponentially associated with the population density, we have linearized the relationship by taking the logarithm of the number of cases.

Another potentially pertinent relationship could be the type

of county (e.g. on an ordinal scale, how urban or rural is it?) vs the number of confirmed COVID-19 cases. Through the federal Rural-Urban Continuum Code (2013), it is possible to quantify this quantity on an ordinal scale from 1 to 8, with 1 specifying the most urban environment and 8 specifying the most rural environment. The relationship between this ordinal scale and the number of cases of COVID-19 is analyzed further.

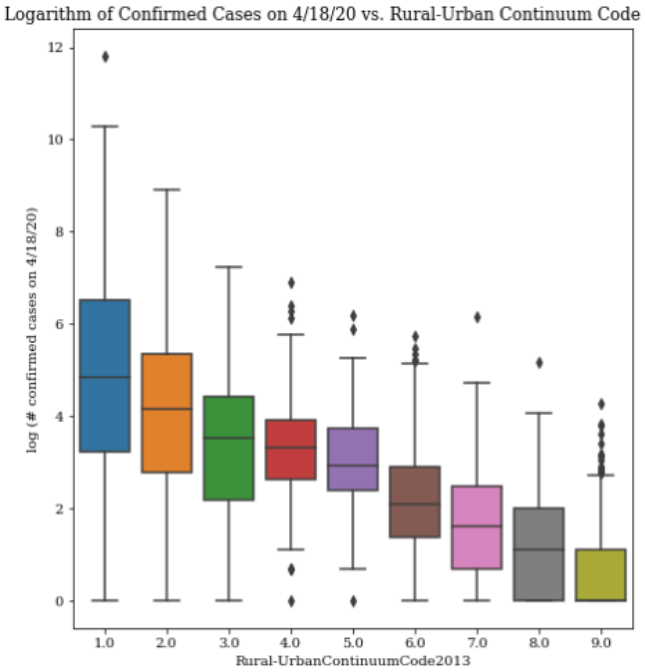

Fig. 2: The natural logarithm of the number of COVID-19 cases as of 04/18/20 are plotted against the Rural-Urban Continuum Code (2013) for each county, showing that the prevalence of COVID-19 monotonically and nearly linearly decreases as a county is classified as more rural. As the number of cases are again exponentially correlated, we have "linearized" the relationship by taking the logarithm of the number of cases for each discrete value.

However, more pertinently, we can examine the relationship between the population density and the Rural-Urban Continuum Code since intuitively, the latter was derived from the former at least to some extent. By quantifying the exact relationship between the two, redundancies and collinear features can be discovered (which won't be useful in this prediction task).

Besides demographic and geographic features, a medical feature that could be helpful in predicting the spread of COVID-19 may be studying pre-existing conditions, which may weaken people's immunoresponses, thereby allowing the virus to spread amongst the population rapidly. One such condition that can weaken the immune system is diabetes, and as such, the relationship between the *growth* in COVID-19 cases with the percent diabetics in each county is examined.

As the virus grows in magnitude, aside from these predictors, many new policy changes have been enacted, including the stay at home orders, the prohibition of

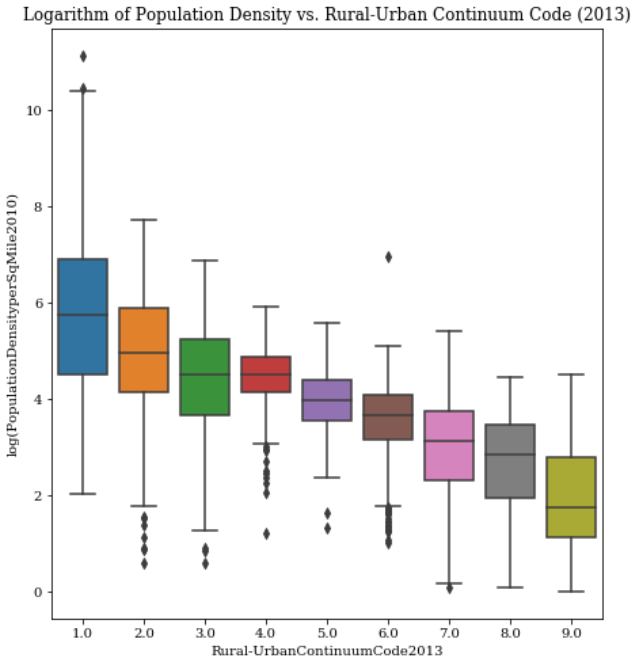

Fig. 3: The natural logarithm of the population density is plotted against the Rural-Urban Continuum Code (2013) for each county, showing that there is a clear exponential association between the two features. We have "linearized" the relationship by taking the logarithm of the population density for each discrete code.

public gatherings above a certain size, closing of certain schools. As before, the most crucial of these features which prevents human-to-human contact largely responsible for virus transmission is the stay-at-home order. The effect of this policy change as it has affected the number of cases is examined.

Finally, we round off the exploratory data analysis by examining a combination of a medical and political feature, which is the number of people on Medicare in each county. The proportion of people that are insured by Medicare in a particular county would be a clear indicator as to each county's ease of access to healthcare and thereby better access to masks, medication, and access to accurate diagnosis through affordable healthcare. Thus, intuitively, this metric ought to be associated in some fashion with the number of confirmed COVID-19 cases, which is examined in a plot.

III. Analysis of Growth Patterns through PCA

Dimensionality reduction through principal component analysis can reveal trends in the growth of COVID-19 cases in various counties of the United States; in other words, provided a multidimensional matrix of day-to-day confirmed case information for COVID-19, we can decipher the importance of each of the principal components that make up the data. Using dimensionality reduction, we are able to quantify the self-contained redundance of some of the dimensions of the data, which in this case are data for cases on a particular day. Simply put, this will allow us to answer the following question: how many days of data of

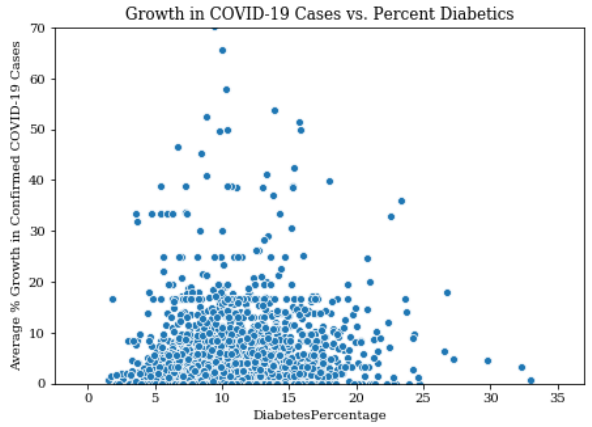

Fig. 4: The average growth in the number of cases over 5 days (till 04/18/20) is plotted against the percent diabetics in each county. Since all the data points appeared to be clustered mostly in the interval [5, 20] x [0, 15] with no visible pattern, it is clear that the features are uncorrelated.

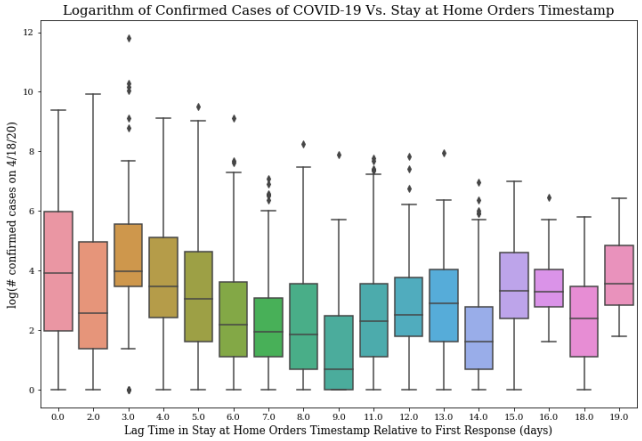

Fig. 5: The natural logarithm of the confirmed COVID-19 cases as of 04/18/20 is plotted against the lag time in enacting stay-at-home orders (starting at day 0, when the first county declared it) for each county. There is a clear non-monotonic curve in the trend, so as to suggest that that the counties who put a stay-in-home provision in place extremely early had already been experiencing exponential growth in cases (e.g. they had been struck much earlier and harder by the virus), and that counties who delayed far too much in putting the stay-in-home provisions suffered as well in terms of the number of confirmed cases. Counties which weren't struck as hard and that did not delay too much didn't suffer as much, as seen around the middle.

COVID-19 cases do we need to explain the variance in its day-to-day spread?

To accomplish this, we use a time series containing day-to-day case data from January to April but focus mostly on data from March onwards since data from January is quite sparse, which makes for less interesting data analysis. Moreover, making local predictions using recent data is more pertinent to the current situation than modeling using older data. We primarily work with data from March 29 till April 18, using singular value decomposition to obtain the principal components matrix.

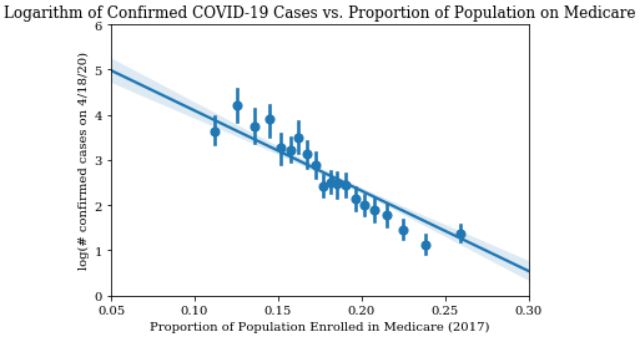

Fig. 6: The natural logarithm of the confirmed COVID-19 cases as of 04/18/20 is plotted against discretized values of the proportion of the population on Medicare (to avoid overplotting), showing a decreasing monotonic and exponential relationship between the number of COVID-19 cases and the proportion of Medicare clients in a county. We have "linearized" the relationship by taking the logarithm of the COVID-19 cases yet again.

Percent Variance Explained by PCs	
Principal Component	Variance Explained (%)
1	99.801977%
2	0.158062%
3	0.027684%
4	0.005049%
5	0.002331%
6	0.001312%

TABLE I: The corresponding percentage of the total variance of the data explained by the principal component is displayed for the first 6 principal components, which make up nearly all the variability in the data. Upon inspection, it is clear that the first principal component is the main contributor to the variability in the data and that it singlehandedly can explain nearly 100% of the variance. The rest of the PCs simply explain the remaining 0.2% of the variance.

With the principal components, it's useful to examine the proportion of the variance in the data explained in each to judge how important they are in determining the distribution of the data. We analyze the principal components matrix and the singular values by displaying the variance explained as a percent of the total variance as well as a (uninteresting, yet surprising) scree plot.

Since we have discovered that simply one component explains nearly 99.8% of the variance, we assert that with a very low-rank approximation of the data, we could very accurately generate the day-by-day COVID-19 case count for each county. By grabbing just the first 3 columns of the matrix U, the first 3 columns and rows of the diagonal matrix Σ, and finally the first 3 rows of V^T, we simply matrix multiply them together to generate a rank-3 approximation as given by the following equation:

$$X_3 = (U_{ij})_{1\leq j\leq 3} \times (\Sigma_{ij})_{\substack{1\leq i\leq 3 \\ 1\leq j\leq 3}} \times (V^T_{ij})_{1\leq i\leq 3}$$

Discussion: This approximation yields simply an approximation of the actual data, but in fact, when

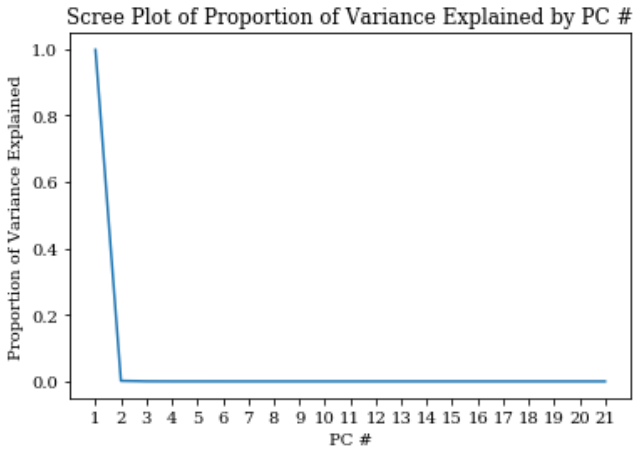

Fig. 7: The scree plot visualizes the proportion of the variance explained by a particular principal component (in this case, for a particular day's worth of COVID-19 case counts) as compared to the total variance of the dataset. Unfortunately, since nearly all the variance is explained by just the first principal component, the scree plot is not quite that interesting.

compared with the actual data, the median absolute error is simply 1 case (we use the median since some exponential outliers in counties in places like New York influence the mean). In fact, the vast majority of the absolute error between the data and prediction hovers between 0 and 5, showing that with simply a linear combination of 3 vectors of case data (for 3 days), we can generate nearly a month worth of case data.

Ultimately, the approximation serves to show that while obviously the case data vectors themselves need to be codependent, they are in fact nearly all reconstructible with simply 3 days worth of information with a reasonable degree of accuracy of being off by 1 death on average for the vast majority of counties. This is quite surprising since we do not expect the data to be linear combinations of other days' data since we would intuitively expect these cases to grow exponentially. Some potential explanations could be that in a more narrow timeframe, exponential growth could in fact resemble linear growth with a higher constant factor (gradient); in other words, with future data, this approach could fail for larger inputs.

In fact, for some particular counties or regions such as New York or San Fransisco, this approximation could be disastrously inaccurate since their behavior is unlike most other counties in terms of linear growth (e.g. their "constant" factor is far higher than most counties). In fact, by inspection, while many of the errors hover around 1-5, there are some errors that are in the 100s or even 1000s, which could indicate a cause for concern in the usage of PCA in confidently predicting information about larger counties. However, for smaller counties, it is clear that we can explain nearly all the variance in the data with very few dimensions.

IV. Regularization and Linear Regression

Since very few days worth of case data can reliably predict the number of cases in any county on any day within the month, we simply focus on our efforts on training a model which can reliably predict the number of cases on one particular day. With that data and the conclusion from PCA, we could determine all the other case information quite easily.

To perform linear regression on the dataset, the predictors (X) are the information about the individual counties (e.g. population density, percent diabetics) and the predicted values (y) are the *natural logarithm* of the number of cases on the chosen date, which is 04/18/20 (the latest data which is available). Given the conclusions from our explanatory data analysis, there are many features which are exponentially correlated with the number of cases, so to apply linear regression, we predict the logarithm of the number of cases instead of the variable itself.

Some assumptions that are made are that each observation for X are independent since the data are scraped from inspections that were conducted at different times in different places. Eventually, we will also show that the errors once the model is generated are normally distributed, which satisfies the normality of errors condition for regression. We assume that the errors are independent since all the pieces of data were, again, obtained independently. Finally, through previous visualizations and data analyses, we can say that for most of the features, there is an exponential correlation with the number of cases, so they will be linearly correlated with the logarithm of the number of cases, which is our y. This condition is satisfied as well.

However, we first apply a simple linear regression (OLS) after the train-test split with X being all the initial features (some are removed since they are not relevant and some are one-hot encoded since they are ordinal) for simplicity. As normal, null values are replaced with the mean for the appropriate feature, and after these data cleaning procedures, there are no remaining categorical data. The RMSE is around 1.04 and the coefficient of determination is around 0.72. While this is certainly acceptable, regularization and some prudent feature selection more would likely improve the cross-validation RMSE, which is a more accurate predictor of test error.

We apply 8-fold cross validation to the data and find that the sum of the cross validation RMSEs is around 9.37. To seek to improve that, we remove some collinear or identical features such as the federal guidelines ban, which is clearly identical for every county since it is at the federal level, some columns with far too many missing values to be useful, and some more metrics not related to the prevalence of COVID-19 (e.g. longitude, latitude). In addition, we regularize using Ridge regression and experiment with various hyperparameters to find the most suitable for this particular prediction task. *Note:* We try standardizing the data, but in fact, that increased both the cross-validation RMSE as well as the training RMSE as a whole, and in

general, the standardized data was **not** well behaved (e.g. often gave RMSEs in the millions when regularized even with very small α values).

However, despite these techniques, the sum of the cross-validation RMSEs did not improve by a significant amount (perhaps around 1% at best), so a randomized algorithm to prune features at random was developed to determine which set of features were the most pertinent to this prediction task. Over several thousand runs, anywhere from 5 to 20 features were pruned and the hyperparameters were then retuned. Through this, the sum of the CV errors dropped by around 5%, which may seem quite low, but in the context of the predicted values being the *logarithm* of the actual number of cases, this is around a 30-40% improvement in the cross-validation RMSE of the actual number of cases.

Some of the pruned features include the timestamp when gatherings over 500 were banned in each county, which were likely thrown out since it was very similar to the feature of when gatherings over 50 were banned or when the stay-at-home order was put into place. Similarly, the timestamp at which public schools closed, restaurants stopped dining-in, and entertainment/gym was restricted were likely codependent (or linearly dependent vectors essentially), and as such, they were removed by the algorithm. A surprising removed feature is the population of older citizens.

To make sure that the final regularized linear model is appropriate, we examine the residuals plot to look for normality (normally distributed residuals imply that a linear fit is appropriate - usually verified through a normal probability plot of residuals).

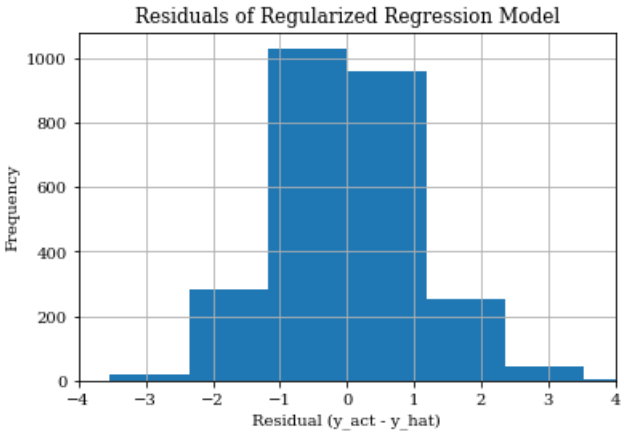

Fig. 8: The histogram of the residuals plot is nearly normal as it is both symmetric about the mean and resembles the classic bell shape of a Gaussian distribution. This implies that the linear fit is appropriate since one of the conditions for using linear regression is that the errors are nearly normally distributed.

Discussion: We also examine the coefficients of the regularized Ridge regression model to determine how many features were used in a meaningful capacity. As expected, the majority of the coefficients in the matrix are nearly zeroed, implying that many of the features were simply not helpful in predictions and led to overfitting.

Finally, the test error is examined after the model is finalized to verify that there isn't a gross error in the prediction task. Fortunately, the test error comes out to around 1.05, which is quite similar to the training set error, around 1.04. When using the model in practice (e.g. for real life applications), we would now train the model on both the train and test set.

With a reasonably good predictor of the number of cases for this particular day (e.g. with quite a decent test error and coefficient of determination 0.72), we are able to extend this result using our conclusion in the PCA section to any other day. Again, as previously discussed, this model has some drawbacks and similarly to PCA, its predictions suffer in quality for counties in which a very strong exponential trend was seen quite early on (e.g. New York). These skew the RMSE, and occasionally, there are some extremely high errors when the errors are transported out of logscale.

Since all the errors are measured in the units of the logarithm of the number of cases, it can be potentially misleading to say that the model was only off by around 1 confirmed case for the most part. While there is definitely a linear trend, it is not as scalable and as accurate as it could be should a more sophisticated non-linear regression model or machine learning be applied in future research. In fact, cross-evaluation with the units being in terms of the actual number of cases proved difficult and improvement through regularization didn't occur at all, so the error units were kept in terms of the logarithm. Additionally, some of the assumptions made for regression may not hold true (e.g. the errors may not necessarily be randomly distributed since there are such outliers - in fact, the residuals histogram shows that there are nearly 2x as many extreme positive residuals as negative, indicating that the model struggled with predicting outliers such as New York).

However, despite this, the linear predictor is quite accurate for most inputs similar to the PCA results, and the coefficient of determination shows that nearly 72% of the variance in the case data is explained by the predictors or the features (X), which is impressive. Combined with the PCA results, the regression model not only becomes stronger as a predictor, but it also delivers a set of useful and redundant features for use in future predictive tasks.

V. Ethical Implications

With the media buzz about the prevalence and growth of COVID-19, it is important for data analyses to stay as unbiased and accurate as possible by not straying from the facts. Transparency in the age of easily manipulable data through either just statistical trickery or through malicious data modification is quite difficult to maintain, and especially with counties wishing to preserve their reputation, they may underreport or misreport their data intentionally, skewing research in strange ways.

Some ways to address these concerns are of course to question the data and provide reasonable explanations for clear trends and strange outliers as well. Additionally, we could cross reference data from different sources and

cross-verify that all of them are reasonably close to one another. While all of these are still fallible in some way, it helps verify that the data is in fact as it should be and not underreported in some way.

One final concern is, of course, causing mass panic through misunderstanding or misreporting findings from this project or others through the media. Fake news is rampant throughout the globe, and the prime subject for people to jump to conclusions is COVID-19 and its growth in the United States. Whether it's politically motivated or otherwise, people are often not selective and thorough in sourcing and verifying information that is delivered to them through social media, news websites, and emails.

Again, a possible way to address this concern is simply verifying information by grabbing info from multiple sources and using some amount of common sense as well.

VI. Summary of Results and Conclusion

Through initial data analysis, dimensionality reduction with PCA, and finally regression, it is clear that the task of predicting the prevalence and growth of COVID-19 comes with many redundancies. From needing only around 3 vectors worth of confirmed cases to generate the entire dataset of around a month with reasonable accuracy to pruning unnecessary collinear or correlated features through regularization, we observe that some surprising features are redundant in this prediction task. Through PCA, it is observed that nearly 99.8% of the variance can be explained by simply utilizing one principal component, and thus, a low rank approximation of the data serves to explain how a full set of data can be obtained just using a calculated linear combination of 3 vectors worth of information.

Furthermore, it is surprisingly discovered through a linear regression model that some age-related features (e.g. proportion of senior citizens) and a few pre-existing conditions (e.g. diabetes, heart disease mortality) were not nearly as associated with the number of cases as other features such as the population density despite intuitively being related. From the initial data analysis, we derive some interesting relationships from the lag time in the stay-at-home order enactment, the Rural-Urban Continuum Code and the Medicare usage as they affect the number of cases.

Overall, through this process, we have obtained a relatively accurate way of not only predicting a given day's number of COVID-19 cases for a particular county through regression with a test RMSE of 1.05, but by harnessing the power of PCA, we know that we are able to reconstruct the data for the days in between with relatively accurate results. In future, it may be useful to construct a more sophisticated and adaptable model which scales better to "outliers" in the data (e.g. this model struggles with predicting the number of cases in New York effectively, and due to the linear nature of the model, it's difficult to remedy that through feature engineering), but for the purposes of serving a general estimate, the linear model suffices.