On the Relation of Country Populations

Devarsh Jhaveri, Mason Rumuly, and Ali Shafi

Abstract

This project explores a variety of feature-selection and machine-learning techniques using World Bank population data. We use LASSO L1 regularized least-squares regression to choose a 5-sparse linear combination of countries to predict population for any other given country. Finally we use the results to draw insight into the ability of a small number of large-population countries to embody information which predicts other populations.

I. Introduction

ONE of the major goals of machine learning is to gain insight into the working of the world. This challenge focuses on the population of a diverse set of countries. Intuitively if one country's population is a good predictor of another country's population over time, it may indicate some useful relation in the influences of certain common factors.
To that end, we have used regression and regularization techniques to identify the best prediction for each country's population using only a linear combination of five other countries. We then attempt to derive useful insight from these results.

October 2, 2018

II. Methods

The algorithmic goal is for each country in the data-set to find a 5-sparse linear combination of other countries which closely predicts the population of that target country by minimizing the Mean Squared Error from the prediction to the true value.
More precisely, where the population of country i in year y is $P_i(y)$, we wish to compute a prediction of the population $\hat{P}_i(y)$ using a maximum of 5 non-zero coefficients. The countries with non-zero coefficients for a given target country are referred to here as constituent countries.

$$\hat{P}_i(y) = \sum_{j \neq i} a_{ij} P_j(y)$$

$$\sum_a ||a_{ij}||_0 \leq 5$$

A. Data Set

We used a sanitized World Bank population data-set for 258 countries and country-aggregates for the years 1960 to 2016. The country-aggregates were removed in order to produce the analyzed in this report.
The years 1960-1999 constituted the training and validation set. The remaining years, 2000-2016, constituted the test set.

B. Exploration

We explored multiple methods for choosing the countries for combination and for carrying out the final regression and prediction. These were validated using k-cross-validation on the training-set with $k = 4$.

1) Greedy Correlation Matrix: The first and most obvious method we explored was to compute the correlation between the target country and the other countries and choose the five with the maximum absolute value correlation coefficient. This strategy intuitively produces the most redundancy in information used in the linear combination.

2) Graham-Schmidt Inspired Correlation Matrix: This method (referred to here as GS) considers the problem as a choice of basis, where each of the countries chosen should embody as many distinct parts of the information as possible. Using the intuition that distinct information should show up as orthogonal elements in the data, we were inspired by the Graham-Schmidt process of choosing an orthonormal basis.
In the resulting algorithm, we computed the correlation between the target country and all other countries, chose the one with the greatest absolute value, and subtracted the projection on to this country's trace from every other country's trace (including the target country) to compute the new correlation. Intuitively, this would remove information embodied in the chosen country from each other country. This process was applied iteratively until 5 countries were selected.

3) LASSO: The final method of choosing countries uses L1-regularized least-squares regression named LASSO (Least Absolute Shrinkage and Selection Operator). This regularization naturally biases its weight vector toward sparsity; an n-sparse vector may be produced by choosing the appropriate hyperparameter α. Since increasing α monotonically decreases the number of non-zero terms in the resulting weight vector, this selection may be efficiently carried out by binary estimation.

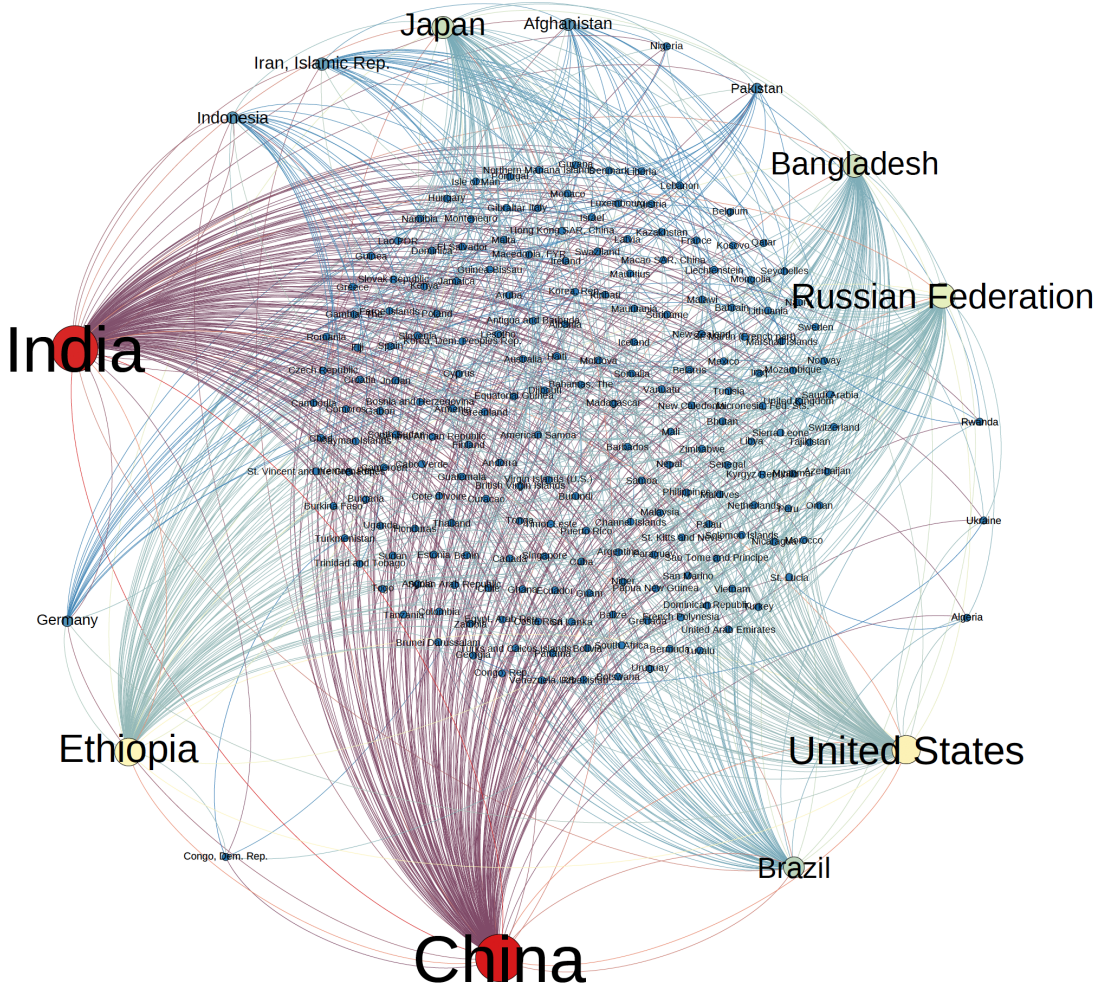

Fig. 1. Network of population relation between countries

4) Centering and Normalization: We also experimented with centering and normalizing the training data before executing each of the methods for choosing countries above.

5) Regression and Prediction: Finally, the prediction was carried out in three ways. Vanilla linear regression (which became possible as the system was no longer over-specified), Ridge Regression (which used L2-regularization to balance the use of each constituent country), and the raw results of LASSO regression (available where that method was applied).

C. Testing and Final Analysis

The two algorithms which performed best in cross-validation for selecting the constituent countries were used for prediction in the test set. The one which performed best out of these was used for analysis.

III. Results

The two best average MSE in cross-validation was achieved by using GS or LASSO to choose the countries. Both performed best with un-normalized and un-centered training data followed by regression using vanilla Least Squares.
On the test set, the LASSO algorithm was the most successful method of choosing the constituent countries. Thus we choose the results from that algorithm for our analysis.

IV. Analysis

For inference to wider patterns from these results, we will consider the results of the regression as a directed graph from constituent countries to the countries of which they are good predictors; that is, where the coefficient in the linear combination was chosen to be non-zero. We ignore the magnitude of these coefficients otherwise, as they relate more to the ratio of population magnitude than to the goodness of prediction. With this abstraction we constructed the directed graph shown in

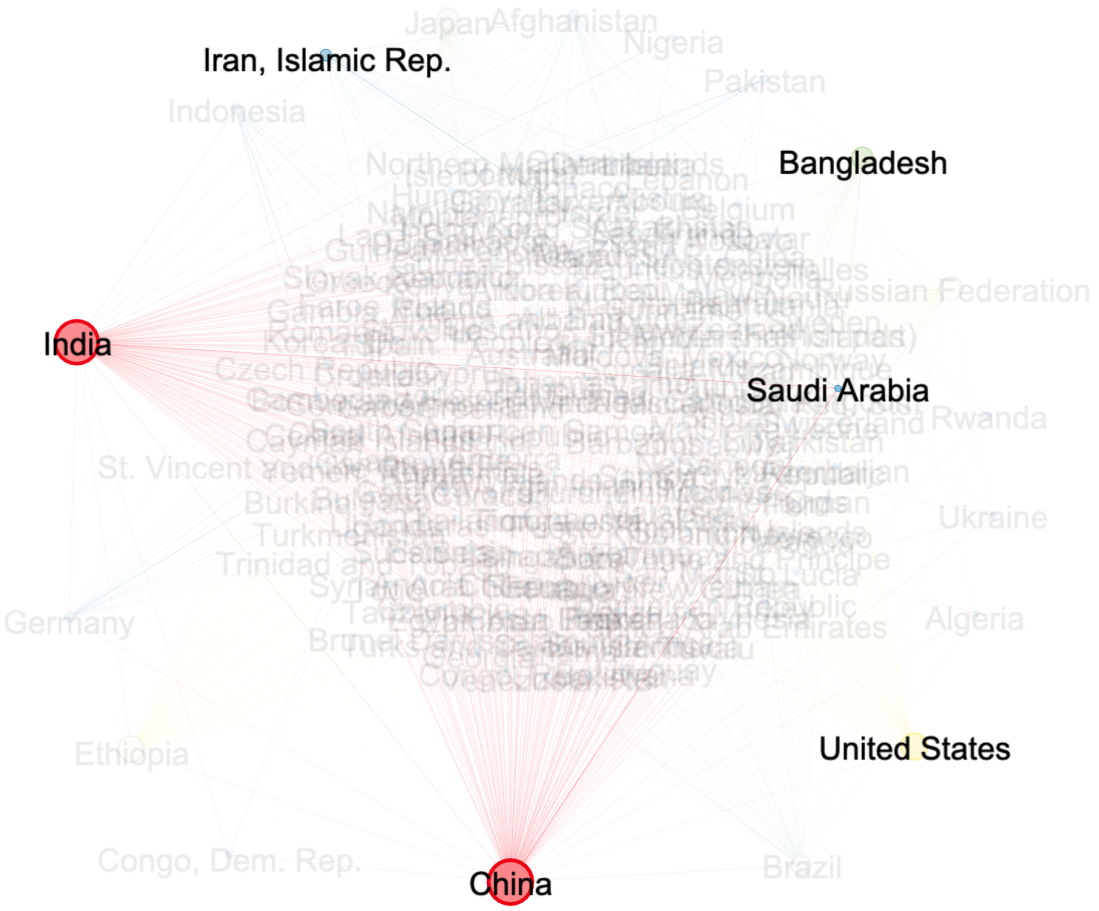

Fig. 2. Network emphasizing countries with many outbound edges

Figure 1.

The outer loop of Figure 1 shows the countries that have outbound edges, that is, are constituents to estimate some countries. These countries are the best predictors of population of the remaining countries, and are only a small subset of all the countries. Countries in the central cloud were never chosen as a predictor of any other country. The outer ring are a diverse set ranging from First World countries like USA and Japan to Third World countries like Afghanistan and Ethiopia. This diverse set of countries embody enough distinct information to predict the population of any given country which incorporates similar demographics and geopolitical information. The information is what allows any given country's population to be predicted.

Applying this insight to a specific example, Saudi Arabia, one may find factors which specify Saudi Arabia's population in its constituent countries USA, China, Bangladesh, India, and Iran (emphasized in Figure 2). Specific factors might be related to a single constituent country, or may be redundantly embodied in multiple to various extents. This analysis does not attempt to infer the specific factor linking each country, but the constituent countries are diverse enough to capture all the information necessary to closely predict the population of Saudi Arabia.

Notice also that these frequently-chosen countries tend to be large, among the largest populations in the world. Aside from regularization bias toward smaller coefficients, large population countries likely are preferable to smaller population countries which embody the same information; due to the Law of Large Numbers, small countries experience larger noise from anomalies than large countries, making the large countries more pure representations of their underlying information. It also implies that a country's size does not significantly affect its rate of population change as a proportion of its population magnitude. Therefore, results from policies or studies done in large countries may be generalizable to countries similar in the respect being studied.

V. Conclusion

LASSO is useful for sparse feature selection which balances diversity of information with redundancy of information, though final prediction is best done with vanilla least squares on the selected features.

Country populations can be well predicted from a subset of countries which consists mostly of the largest countries, indicating that there is a small set of common information which large nations are diverse enough to capture and apply regardless of country population. Future work may seek to find the specific factors the existence of which are implied by these efficient interrelations.