A statistical analysis on factors that contributed/slowed down the spread of COVID-19

Project for the exam: Machine learning, statistical learning, deep learning and artificial intelligence

Marzio De Corato

September 1, 2020

"Text"

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

1 INTRODUCTION

The recent pandemic diffusion of the virus *SARS-CoV-2* connected to the *COronaVIrus Disease 19* raised up a scientific issue about the different spread ratio among the different population clusters such as cities,regions and states. This behaviour can be ascribed to the fact that the features of these clusters, such as the densities, the public transports as well as the individual mean incomes are very heterogeneous [3, 4]. Furthermore the individual policies undertaken by the decision maker of these clusters have a large effect on the spread of the virus [5]. A good tool for an analysis about the correlations of the spread of the virus with the cluster feature is the Principal Component Analysis (PCA): this approach allows to visualize in to a single 2D plot the main statistical correlation of a data set. It is worth nothing that this type of analysis represent only the first step for the identification of causal correlation between the features: the findings obtained via PCA should be validated with a model; however this second step is out from the aim of this work. In this work three different type of cluster were considered: the Italian provinces (*province*), the Italian regions (*regioni*), and 150 countries. The choice of the of Italian provinces was motivated by the fact that this the smallest cluster for which the daily cases of COVID19 are easily available, on the other hand the regions provide aggregate data (in particular about the sanitary system) that are not easily retrievable for the provinces. Finally the inclusion of the states is aimed to see if the statistical correlation obtained for the smaller cluster were still valid in a macroscopic scale. For this last cluster the lock-down policy and its effect was not considered, because the data collection of the policies for each 150 States was too time consuming, however this analysis may be an outlook for a future update of this work.

2 DATA DESCRIPTION

In this section a brief overview about the data analysed is provided: in particular, for each cluster,a brief description is provided about the data gathering and assembling.

2.1 Provinces

For this cluster the cumulative number of cases of COVID19 up to a certain date was obtained from the github website of *Protezione Civile* [6]. The overall number of total cases was than divided (normalized) by the population number of its province. The population number,

up to 2019, was retrieved from Istituto Nazionale di Statistica (ISTAT) [7]. The size of the province was also considered in order to evaluate the density. Since the diffusion of a virus is causally correlated to connectivity and public mobility services [9, 8] we considered also a set of economic indices in order to evaluate the industrialization and the wealth of each province: these are the mean income for person, the public transport, the pollution and the unemployment. The first one was taken from *Ministero dell'Economia e delle Finanze* as reported in the following website [10] while the other ones were obtained from the ISTAT databases [7].

3 THEORETICAL BACKGROUND

As pointed out by Mackay [1] the basic idea of unsupervised learning is to mime human behaviour is to find regularities in data and group them. Among the different techniques of unsupervised learning [2] here we considered the PCA: in this section the basic ideas of this tool will be focused following the approach proposed by James et al. [2]. The aim of PCA is to plot n observations with p features in only one 2D plot,with the least possible loss of information, instead of $\binom{p}{2}$ 2D plots. To do so the PCA chooses, among the possible axis formed by the linear normalized combination of the features, the two associated with the largest variance, given that these two axis are orthogonal. Thus, supposing that the means of the features are null, the problem is to find the coefficients (usually called loadings) ϕ_{ij} that solves the following optimization problem:

$$max \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \tag{1}$$

with the constraint

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1 \tag{2}$$

The loadings define the principal components (that have to be orthogonal) :

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + ... + \phi_{p1} x_{ip} \tag{3}$$

If the all the principal components are taken in to account no information is lost, however the complexity of it due to high dimensionality is the same of the starting one; however if we sacrifice the p-2 component, that have a lower variance with respect to the first two, the complexity is dramatically reduced to a 2D plot. In this case the

toll paid is the lost of the information contained in the $p-2$ component. From a computational point of view this task can be performed with the standard techniques used for solving a eigenvalue problem. Once the two principal component are found, the data can plotted into the new coordinate set; the advantage of this new representation lies on the fact that if the first two principal component loading vectors are plotted a bird's-eye view of the statistical correlation between the features is obtained: basically the cosine of the angle between the loadings approximates the statistical correlation, while the position of the data with respect to the loadings will plot its feature. In this way an intuitive representation of the data, their features and the correlation among them is provided. Finally this can be accompanied by a plot of the correlation matrix as it is done in this work.

REFERENCES

[1] David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[3] Abiel Sebhatu, Karl Wennberg, Stefan Arora-Jonsson, and Staffan I Lindberg. Explaining the homogeneous diffusion of covid-19 nonpharmaceutical interventions across heterogeneous countries. *Proceedings of the National Academy of Sciences*, 2020.

[4] Piotr Skórka, Beata Grzywacz, Dawid Moroń, and Magdalena Lenda. The macroecology of the covid-19 pandemic in the anthropocene. *PloS one*, 15(7):e0236856, 2020.

[5] Per Block, Marion Hoffman, Isabel J Raabe, Jennifer Beam Dowd, Charles Rahal, Ridhi Kashyap, and Melinda C Mills. Social network-based distancing strategies to flatten the covid-19 curve in a post-lockdown world. *Nature Human Behaviour*, pages 1–9, 2020.

[6] Protezione Civile. https://github.com/pcm-dpc/COVID-19.

[7] Istituto nazionale di Statistica (ISTAT). https://www.istat.it/it/dati-analisi-e-prodotti/banche-dati/statbase.

[8] Moritz UG Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Louis Du Plessis, Nuno R Faria, Ruoran Li, William P Hanage, John S. Brownstein, Maylis Layan, Alessandro Vespignani, Huaiyu Tian, Christopher

Dye, Oliver G. Pybus, and Samuel V. Scarpino. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020.

[9] Alun L Lloyd and Robert M May. How viruses spread among computers and people. *Science*, 292(5520):1316–1317, 2001.

[10] Ministero dell'Economia e delle Finanze. https://www.idealista.it/news/finanza/economia/2019/04/02/130615-la-mappa-del-reddito-pro-capite-nelle-province-italiane.