COVID-19 : Control Measures' Importance Based on Non-negative Matrix Factorization

Lukman O. Olagoke
adderbyte@icloud.com

Ahmet E. Topcu
College of Engineering and Technology
American University of the Middle East, Kuwait
ahmet.topcu@aum.edu.kw

ABSTRACT

COVID-19 was declared a global pandemic, and several control measures were put in place to mitigate its spread. It is clearly important that we learn about the effectiveness of control measures across different regions. Such knowledge will help both governments and caregivers understand the impact of the control measures and inform good decision making. It is also clearly desirable to reinforce effective control measures rather than rolling out numerous measures that will not have the right impact. In this work, we develop a model to predict the performance impact of control measures. Since the available data are very limited, we use non-negative matrix factorization to achieve a sparse representation of the input feature space. Furthermore, we rank feature importance using the gradient boosting method (LightGBM) and the regularized linear model (ElasticNet). Lastly, we compare the importance of ranking for high performing regions against regions that are still lagging in performance. It is hoped that such control measure analysis will help policy makers understand what measures are effective and essential in controlling the pandemic.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

COVID-19, Non-negative matrix factorization, feature selection, LightGBM, lasso regularization

ACM Reference Format:
Lukman O. Olagoke and Ahmet E. Topcu. 2018. COVID-19 : Control Measures' Importance Based on Non-negative Matrix Factorization. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Coronavirus virus disease 2019 (COVID-19) was discovered in Wuhan in 2019 [16]. It was declared a global pandemic by the World Health Organization [18]. Thereafter, the governments of different countries initiated different measures to curb its spread. A dataset containing details about the control measures implemented by different countries can be found in [1]. This paper was motivated by the need for an explainable approach to rank the importance of control measures in mitigating the spread of this disease. This might help policy makers understand which control measures are effective, for example, and hence should be given preference over others.

Linear dimensionality reduction methods are very useful for the analysis of high-dimensional data [4]. This is because they can generate a low-dimensional linear mapping of the data while preserving some feature of interest [4]. A review of linear dimensionality reduction methods has been provided in [4]. Of importance to us in this work is the non-negative matrix factorization (NMF) method [15]. An attractive feature of non-negative matrix factorization is that it helps extract sparse and meaningful features from non-negative high-dimensional data [10]. This is particularly important for this present work since we desire that the results be generally interpretable and understandable.

Furthermore, LightGBM [11] is a gradient boosting decision tree algorithm that uses *gradient-based one side sampling* and *exclusive feature bundling* (EFB) [11]. EFB explores the sparsity of features in higher-dimensional space to speed up learning by bundling mutually exclusive features into what is called an "exclusive feature bundle" [11]. Generally, all tree boost-based methods are invariant under monotone transformation of the feature space and are thus less sensitive to outliers [12]. They also exhibit internal feature selection [12], which is useful for feature importance ranking. Thus, we have NMF, which helps extract sparse features, while on the other hand, we use LightGBM, which exploits sparsity to improve learning. The first task in this work consists of proposing a methodology for combining NMF and LightGBM models for the purpose of feature importance selection.

To check the validity of the methodology and results, a variant of the lasso regression model was used as a base model. This model variant consists of an elastic net [13] that uses the ℓ_1 norm and sets ℓ_2 norm to 1 [7]. It is worth noting that the ℓ_1 norm helps to enforce sparse solution vectors [19]. The sparse solutions can be exploited for feature selection [14].

2 BACKGROUND

2.1 Non-negative matrix factorization (NMF)

The aim of NMF is to find the factorization of a real non-negative $(f \times n)$ matrix X into non-negative matrices W and H such that:

$$\underset{f\times n}{\mathrm{X}} \approx \underset{f\times r}{W} \times \underset{r\times n}{H}, \tag{1}$$

Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

where $r < \min(n, m)$ [5].

This is a typical case where we have high-dimensional data with f components and n samples represented in low r dimensions with basis elements w_k, where $1 \leq k \leq r$. For each vector x_j of X, we have:

$$\mathbf{x}_j \approx \sum_{l=1}^{r} h_{(l)j} \mathbf{w}_l, \tag{2}$$

where h_j of H is the coordinate or activation of the corresponding v_j [10]. This implies that each vector of X can be recovered as the linear combination of the bases W with the activations or coordinates given by H.

The heatmap in Figures 1a, 1b, 1c, and 1d provides a visual representation of the approximation that has been carried out on the dataset of control measures. The similarity between the constructed data from matrices W and H is almost invisible by typical visual inspection.

In order to measure the quality, the approximation of a cost function is necessary. For example, the Frobenius norm can be used [5]. A useful review of other techniques for efficient computation of the approximation with their corresponding cost functions is provided in [3].

2.2 Model setup

Usually, we assume that we have training sample data from which we would like to obtain a good estimate $\mathbf{F}(x)$ of the true functional mapping $\mathbf{F}^*(x)$ for all $(\mathbf{x}, y)$ data element pairs. To achieve a meaningful approximation, one usually aims to minimize some loss function over the joint distribution of x and y [9]:

$$\mathbf{F}^* = \arg\min_{\mathbf{F}} \mathbb{E}_{(y,x)} \mathbf{L}(y, \mathbf{F}(\mathbf{x})). \tag{3}$$

To make the estimation tractable over the joint distribution of $(\mathbf{x}, y)$, $\mathbf{F}$ is restricted to a parameterized function approximator with parameter $\Theta \in \mathbb{R}^d$ for some $d \in \mathbb{N}$ [8]. Thus, $\mathbf{F}$ takes the following generalized form:

$$\mathbf{F}(\mathbf{x}; \{\alpha_m, \Theta_m\}_1^m) \tag{4}$$

for some $m \in \mathbb{N}$. Furthermore, we can now minimize the loss function over the parameter space [12]:

$$\{\alpha_m, \Theta_m\}_1^m = \arg\min_{\{\alpha,\Theta\}_1^m} \{\mathbf{L}(y_i, \mathbf{F}(\mathbf{x}_i; \{\alpha_m, \Theta_m\}_1^m))\}_{i=1}^n. \tag{5}$$

Now we make further assumptions about the vectors $\mathbf{x_i}$ of $\mathbf{X}$ (these are the feature vectors). We assume that they have non-negative real values such that $\mathbf{X}$ is a non-negative matrix. This allows for the factorization of $\mathbf{X}$ using NMF. Let us factorize $\mathbf{X}$ using NMF so that we have:

$$\underset{f \times n}{\mathbf{X}} \approx \underset{f \times r}{W} \times \underset{r \times n}{H}. \tag{6}$$

Since W represents the set of basis elements for the representation of X r-dimensional space, we retain W and rewrite the minimization as:

$$\{\alpha_m, \Theta_m\}_1^m = \arg\min_{\{\alpha,\theta\}_1^m} \{\mathbf{L}(y_i, \mathbf{F}(\mathbf{w}_i; \{\alpha_m, \theta_m\}_1^m))\}_{i=1}^n. \tag{7}$$

The aim of doing this is to recover a sparse representation of the original input data. The sparse representation would help force some features to zero [19] and make feature selection efficient. Particularly, we stress that, in the case of COVID-19, there are not enough datasets per country for efficient feature importance learning, and therefore, a sparse representation would be very useful. In this paper, **F** will be either LightGBM [11] or adapted ElasticNet [13].

3 EXPERIMENT

The initial motivation for the methodology so far described has been the need to rank the importance or impact of control measures for COVID-19. We wanted to know what the effective control measures are that could minimize the effect of the pandemic. Accordingly, the first important resource that we collected was government COVID-19 mitigation data as provided in [1]. In addition, we collected data about confirmed and recovered COVID-19 cases from [6].

3.1 Problem setup

We computed the performance index for each day per country as:

$$\mathcal{P} = \frac{\Delta R - \Delta C}{\Delta R + \Delta C}, \tag{8}$$

where $\mathcal{P}$ is the performance index, ΔR is the marginal or incremental change in recovered cases, and ΔC is the marginal or daily change in confirmed cases as discussed in [17]. This performance index has been shown to have a negative correlation with the reported cases of death, which is a very important property since the ultimate aim is to limit the cases of death resulting from COVID-19. Furthermore, [17] noted that control measures were important for improvement in performance. The control measures were also important for flattening the epidemic curve, as reported in [2]. We therefore have two problems to solve:

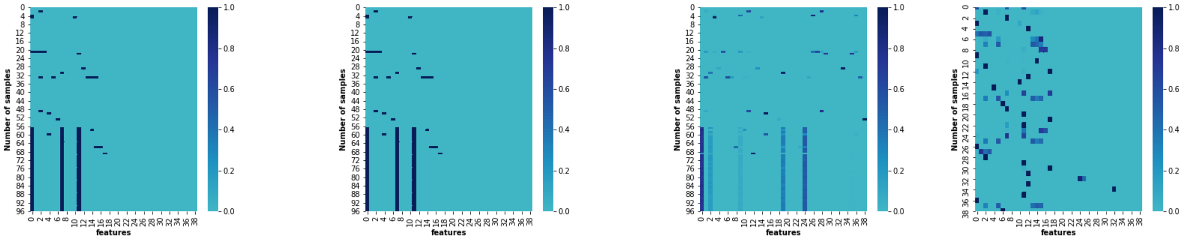
Problem 1. *Given the data on government control measures, build a model to predict the performance index.*

Problem 2. *Given the model, rank the government control measures.*

There were 39 control measures (model features) in the dataset. Since it would be too tedious to build a model for each country, we took 2 countries with good performance indexes and 2 countries with poor performance indexes per region. Afterwards, we analyzed the results obtained.

4 RESULTS

The targets are the performance indexes that have been precomputed from the recovered and confirmed COVID-19 cases. The features are the encoded lockdown information (1 if a particular control measure is applied on a particular day and zero if not). First we compare the performance of the training without applying NMF on the dataset with that obtained when NMF is applied and we compare the results.



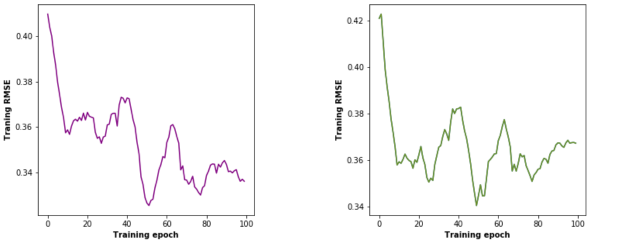
(a) The heat map of original data

(b) Countries across Europe

(c) The "W" matrix (basis elements)

(d) "H" matrix or coordinates

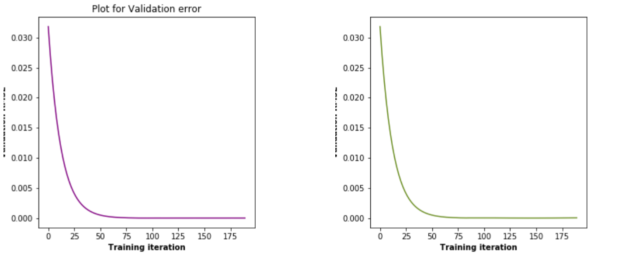
Figure 1: Heatmap of matrices W, H generated using NHF on the data. Note similarity between real heatmap and reconstructed heatmap from W and H. The reconstruction error is of order $1e-5$.



(a) Original feature set as input

(b) Basis matrix "W" of feature set as input

Figure 2: Training error plot.



(a) Original feature set as input

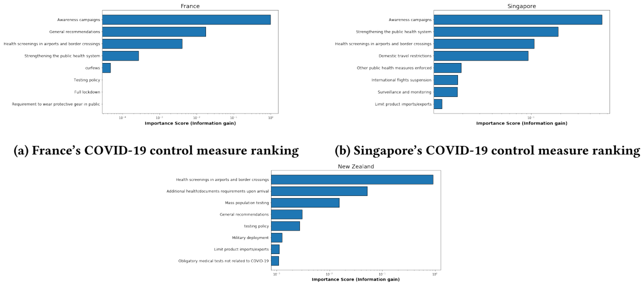
(b) Basis matrix "W" of feature set as input

Figure 3: Validation error plot.

4.1 Good performance

The results for the measure importance for selected countries are shown in Figures 4a, 4b, and 4c. It is obvious that many factors not considered in this work such as culture and economic policies might have an influence on what control measures to implement. However, the underlying pattern as reflected in the figures reflects the need for awareness, strengthening the health system, and perhaps a good control measure to prevent infected people from moving across regions. Countries have been carefully selected here for their good performance in recent times.

For comparison, similar analysis was performed using ElasticNet as shown in Figures 5a, 5b, and 5c. These results show trends similar to those obtained from LightGBM.

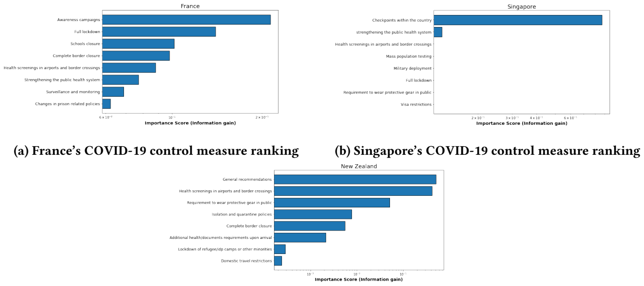


(a) France's COVID-19 control measure ranking

(b) Singapore's COVID-19 control measure ranking

(c) New Zealand's COVID-19 control measure ranking

Figure 4: Control measure importance for countries having good performance.



(a) France's COVID-19 control measure ranking

(b) Singapore's COVID-19 control measure ranking

(c) New Zealand's COVID-19 control measure ranking

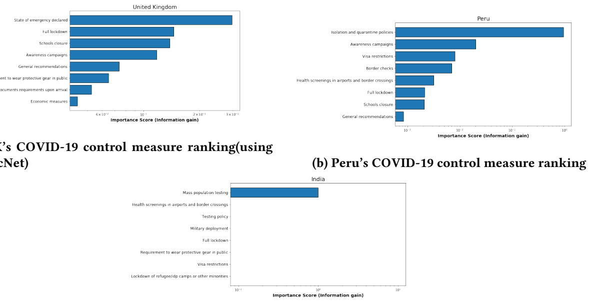
Figure 5: Performance comparison using ElasticNet.

4.2 Poor performance

A small number of countries also showed a lag in performance. The results for some of those countries are presented in Figures 6a, 6b, and 6c. The aim of this step is to check what high performing countries have implemented so that countries that are lagging in performance could also try to implement such measures.

5 DISCUSSION

The best method would be to build a personalized model for a specific country and then compute the feature importance. Such a model would better reflect the underlying measure that might be relevant for that particular country since nations, like individuals, can have differences in behavior. However, should we attempt to undertake such a task, it would entail a huge number of models.



(a) UK's COVID-19 control measure ranking(using ElasticNet)

(b) Peru's COVID-19 control measure ranking

(c) India's COVID-19 control measure ranking

Figure 6: Control measure importance for some countries lagging in performance

What we have done here is build a generic model and test for performance within this same generic framework. Thus, the use of NMF is central to this study since it provides us with sparse and explainable representation.

The general idea and the motivation were to compare what measures high performing countries are implementing that others could apply to achieve the same outcomes. In the end, the aim would be to prevent unnecessary and avoidable casualties in all states. Generally, we could observe a trend of health system strengthening, testing, and checks in appropriate places to avoid movement from infected regions to uninfected regions. There is also an overlap of measures between high performing countries compared to countries that are lagging. In such cases, it would be relevant to check the order and values of the importance scores. It should be emphasized that some of the countries that had high performance indexes were previously performing poorly (i.e., with high numbers of reported cases), so the spikes in performance are explainable in terms of the measures that were rolled out. These results could help countries that have not witnessed high numbers of reported cases to identify the best possible response in the event that cases begin to increase.

We make no assertive claim that the models in this work are bullet-proof; clearly, it can be seen that the quality of the model depends on the data that were released and the authenticity of such data. However, computational methods like the ones described in this work could help put measures into a comparative perspective and thus prevent further disaster. Notably, it would be important that policy makers turn to this kind of quantitative analysis to help make educated decisions that positively affect lives. In the end, it is the safety of the population that matter.

6 CONCLUSION

In this work, we have proposed a methodology and a model to analyze the importance or effectiveness of control measures. Policy makers and caregivers use these measures to alleviate the spread of COVID-19. This has the potential of helping caregivers and policy makers understand the impact of the control measures that have been rolled out. It is essential to apply the most efficient measures from the list of available actions to reduce the impact of the pandemic. This study helps us to understand what standards are helpful for reducing the number of cases and ensuring public confidence.

REFERENCES

[1] ACAPS. May 18, 2020. *COVID19 Government Measures Dataset*. ACAPS. https://www.acaps.org/covid19-government-measures-dataset

[2] Roy M Anderson, Hans Heesterbeek, Don Klinkenberg, and T Déirdre Hollingsworth. 2020. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *The Lancet* 395, 10228 (2020), 931–934.

[3] Benjamin Cauchi. 2011. *Non-negative matrix factorisation applied to auditory scenes classification*. Master's thesis. Université Pierre et Marie Curie, France.

[4] John P Cunningham and Zoubin Ghahramani. 2015. Linear dimensionality reduction: Survey, insights, and generalizations. *The Journal of Machine Learning Research* 16, 1 (2015), 2859–2900.

[5] Lee Daniel D and Seung H Sebastian. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*. 556–562.

[6] Dong Ensheng, Du Hongru, and Gardner Lauren. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* (2020).

[7] Pedregosa F., G. Varoquaux, Gramfort A., V. Michel, Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., A. Passos, D. Cournapeau, Brucher M., Perrot M., and Duchesnay E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[8] Rick Farouni. 2017. A Contemporary Overview of Probabilistic Latent Variable Models. *arXiv preprint arXiv:1706.08137* (2017).

[9] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[10] Gillis and Nicolas. 2014. The why and how of nonnegative matrix factorization. *Regularization, optimization, kernels, and support vector machines* 12, 257 (2014), 257–291.

[11] Ke Guolin, Meng Qi, Finley Thomas, Wang Taifeng, Chen Wei, Ma Weidong, Ye Qiwei, and Liu TieYan. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*. 3146–3154.

[12] Friedman Jerome H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[13] Zou Hui and Hastie Trevor. 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* 67, 2 (2005), 301–320.

[14] Tang Jiliang, Alelyani Salem, and Liu Huan. 2014. Feature selection for classification: A review. *Data classification: Algorithms and applications* (2014), 37.

[15] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788–791.

[16] Zhu Na, Zhang Dingyu, Wang Wenling, Li Xingwang, Yang Bo, Jingdong Song, Zhao Xiang, Huang Baoying, Shi Weifeng, Lu Roujian, et al. 2020. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine* (2020).

[17] Lukman Olagoke and Ahmet Topcu. 2020. *Characterizing the effectiveness of COVID-19 mitigation measures: A data-centric approach*. Under Review, theBMJ. Retrieved May 18, 2020 from https://github.com/adderbyte/covid_19_response

[18] World Health Organisation. 2020. *WHO update Timeline*. WHO. Retrieved May 18, 2020 from https://www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19

[19] Hastie Trevor, Tibshirani Robert, and Wainwright Martin. 2015. *Statistical learning with sparsity: the lasso and generalizations*. CRC press.