Fuzzy Systems and Data Mining IX
A.J. Tallón-Ballesteros and R. Beltrán-Barba (Eds.)

doi:10.3233/FAIA231015

Bayesian Matrix Completion for Ranking COVID-19 Cases

Yushu Cao and Lei Zhang[1]
School of Economics and Management, Beijing Jiaotong University, Beijing, 100044, China

Abstract. Due to the surge in COVID-19 cases, hospitals have had to receive many more patients than before, which has brought unprecedented pressure to the hospital system. Therefore, the emphasis of medical decision-making has shifted from reaching the best treatment effect to prioritizing the treatment of COVID-19 patients by hospitals, which is key to relieving the pressure on the hospital system and reducing the overall mortality rate of COVID-19. There is no doubt that establishing the prioritization of COVID-19 cases is fundamental and pivotal for hospitals to achieve the shift in medical decision-making. Prioritization of COVID-19 cases in previous studies was mostly based on one patient characteristic, mainly including age, health conditions, and gender. This paper focuses on two patient characteristics at the same time. The probability that a COVID-19 patient who died had a given health condition in a given age group is calculated using the matrix completion technique based on the high-rank assumption of Bayesian matrices and the properties of Markov matrices. The calculated results show that doctors should give patients over 55 with respiratory diseases, patients over 65 with circulatory diseases, and patients over 65 with diabetes a higher prioritization in COVID-19 treatment.

Keywords. COVID-19, matrix completion, high-rank assumption, Bayesian matrices, Markov matrices, treatment prioritization

1. Introduction

Since the outbreak of the COVID-19 epidemic, it has seriously influenced people's daily lives and affected various industries and sectors around the world, including healthcare, economy, and society, to varying degrees [1]. According to the World Health Organization, the number of COVID-19 related deaths has reached 744,175 as of January 2, 2023 [2]. However, having COVID-19 is not the only factor that contributes to COVID-19 related deaths. These deaths are also influenced by other factors, including the age and health condition of patients. Specifically, the mortality rate of patients with underlying diseases is four times that of patients without underlying diseases, and patients older than 41 years are considered to be at higher risk [3]. In addition, due to the surge in COVID-19 cases, hospital systems are under unprecedented pressure. In this case, the focus of medical decision-making shifts from achieving the best treatment outcome for individual patients to providing treatment for a larger patient population to ensure the maximum overall benefit of all patients [4]. Therefore, determining the prioritization of COVID-19 cases and maximizing the use of hospital resources to reduce

[1] Corresponding Author: Lei Zhang, School of Economics and Management, Beijing Jiaotong University, Beijing, 100044, China. Email: zhlei@bjtu.edu.cn.

the overall mortality rate of patients with COVID-19 has become the top priority of hospital management. Due to the relatively crude nature of relevant data, previous related research was limited to a single-factor statistical analysis of age and health condition, so the prioritization of patient treatment was also relatively rough, limited to a single-factor priority [3]. Hence, this paper focuses on combining age and health conditions to determine more precise treatment priorities and provide more accurate recommendations to hospital systems.

Low-rank matrix completion techniques are currently being studied in the literature on missing value completion. Trever Hastie et al. developed a brand-new algorithm that can handle data from the Netflix challenge by applying low-rank singular value decomposition (SVD) to improve previous matrix completion algorithms [5]. Xiaofeng Liu et al. took advantage of the high spatial correlation and consistency of the air pollutant spatial matrix and used the low-rank matrix completion algorithm to fill in the missing values of the environmental station data [6]. Xuelong Li et al. introduced a new method for fusing HS and MS images using nonlocal low-rank tensor approximation and sparse representation [7]. Feiping Nie et al. introduced a non-convex regularizer and used it to develop two models for matrix completion based on the low-rank assumption [8].

However, since there is a causal relationship between patient age and health condition, age and health condition form a high-rank Bayesian matrix. Therefore, low-rank matrix completion is no longer applicable. This paper performs Bayesian matrix completion under the assumption of high-rank matrices [9] to provide hospitals with more accurate treatment prioritization of COVID-19 cases during the pandemic.

2. Data and Bayesian Matrix

The data comes from the official website of the US government, and the dataset displays health conditions by age group and jurisdiction of occurrence that were reported in connection with deaths from COVID-19. This paper selects data recorded for the entire United States from January 1, 2020, to January 29, 2023, and properly cleans the data to remove very few records of unknown age. Health conditions recorded in the dataset include respiratory diseases, circulatory diseases, sepsis, malignant neoplasms, diabetes, obesity, Alzheimer's disease, vascular and unspecified dementia, renal failure, intentional and unintentional injury, poisoning, and other adverse events, as well as all other conditions and causes (residual), and COVID-19 (COVID-19 is the only health condition that leads to COVID-19 deaths). There are eight different age groups in the dataset, including 0-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, and 85+. Figure 1 shows the proportion of various health conditions, while Figure 2 shows the proportion of various age groups.

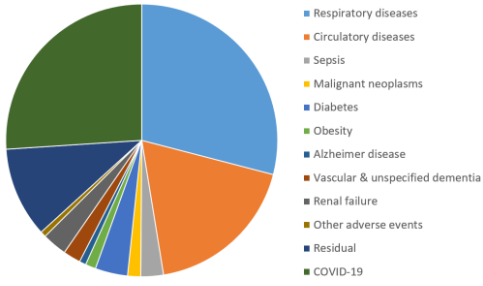

Figure 1. The proportion of various health conditions.

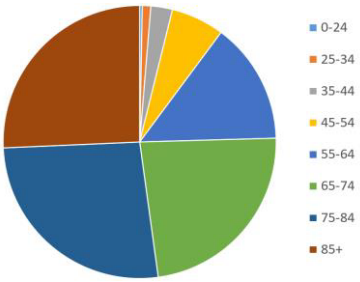

Figure 2. The proportion of various age groups.

Let the health conditions be written as $C = (c_1, c_2, c_3, \cdots, c_n)$, and the age groups be written as $A = (a_1, a_2, a_3, \cdots, a_m)$. The concrete Bayesian matrix B $\big((n+m) \times (n+m)\big)$ can be written as follows, and it is divided into four main parts:

$$B = \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix} = \begin{pmatrix} 0 & P(A|C) \\ P(C|A) & 0 \end{pmatrix} \tag{1}$$

b_1 refers to the association between each age group. It can be seen as 0 since there is no association between different age groups. b_2 refers to the probability of a COVID-19 patient dying due to an i age group, based on a given health condition. b_3 refers to the association between each health condition. It can be seen as 0 since there is almost no association between different health conditions because the dataset has categorized different health conditions. b_4 refers to the probability of a COVID-19 patient dying due to a j health condition, based on a given age group.

3. Bayesian Matrix Learning

In the Bayesian matrix B, b_1 and b_4 are equal to 0, and b_2 and b_3 represent conditional probabilities. Thus, all elements in the matrix are in the closed interval from 0 to 1. Additionally, since $(c_1, c_2, c_3, \cdots, c_n)$ and $(a_1, a_2, a_3, \cdots, a_m)$ are collectively exhaustive events for the entire dataset, the sum of all elements in each column in the matrix is 1. Therefore, according to the characteristics of the Markov matrix, it can be determined that the Bayesian matrix B is a Markov matrix. From the properties of Markov matrices, the Bayesian matrix B has an eigenvalue of 1 [10]. The next task is to find the eigenvector with eigenvalue 1 and find the Bayesian matrix B according to the eigenvector and high-rank assumption.

$$\begin{pmatrix} 0 & P(A|C) \\ P(C|A) & 0 \end{pmatrix} \begin{pmatrix} P(A) \\ P(C) \end{pmatrix} = \begin{pmatrix} P(A) \\ P(C) \end{pmatrix} \tag{2}$$

i.e.

$$\begin{pmatrix} \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} & \begin{pmatrix} P(a_1|c_1) & \cdots & P(a_1|c_n) \\ \vdots & \ddots & \vdots \\ P(a_m|c_1) & \cdots & P(a_m|c_n) \end{pmatrix} \\ \begin{pmatrix} P(c_1|a_1) & \cdots & P(c_1|a_m) \\ \vdots & \ddots & \vdots \\ P(c_n|a_1) & \cdots & P(c_n|a_m) \end{pmatrix} & \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \end{pmatrix} \begin{pmatrix} P(a_1) \\ \vdots \\ P(a_m) \\ P(c_1) \\ \vdots \\ P(c_n) \end{pmatrix} = \begin{pmatrix} P(a_1) \\ \vdots \\ P(a_m) \\ P(c_1) \\ \vdots \\ P(c_n) \end{pmatrix} \quad (3)$$

It is crystal clear that $s = \begin{pmatrix} P(A) \\ P(C) \end{pmatrix}$ is the eigenvector with eigenvalue 1, which is also the principal eigenvector of the Bayesian matrix B. Since B is a Markov matrix, it becomes a stable state after limiting self-multiplication. Therefore, there is:

$$\lim_{n\to\infty} B^n = S = (s, s, \cdots, s) \quad (4)$$

i.e.

$$BS = SB = S \quad (5)$$

And because the principal eigenvector s is known, the task of finding the Bayesian matrix B can be transformed into finding the maximum eigenvalue of the Bayesian matrix B. For eigenvalue λ, there is:

$$Bs = \lambda s \quad (6)$$

$$SBs = \lambda Ss \quad (7)$$

$$s^T SBs = \lambda s^T Ss = 2\lambda s^T s \quad (8)$$

$$\lambda = \frac{s^T SBs}{2s^T s} = \frac{s^T Ss}{2s^T s} \quad (9)$$

Utilizing the characteristics of the Markov matrix as a constraint on the solution, after maximizing the eigenvalue λ, the Bayesian matrix B can be solved, and a more accurate prioritization of COVID-19 case treatment can be obtained.

4. Experiment and Result

The data from the official website of the US government demonstrates the frequencies of each age group and each health condition. Treating frequencies as probabilities, the probability for each age group is: 0-24 (0.31%), 25-34 (1.00%), 35-44 (2.56%), 45-54 (6.30%), 55-64 (14.39%), 65-74 (23.27%), 75-84 (26.48%), 85+ (25.69%). The probability for each health condition is: respiratory diseases (29.04%), circulatory diseases (18.39%), sepsis (2.69%), malignant neoplasms (1.56%), diabetes (3.86%), obesity (1.24%), Alzheimer's disease (0.79%), vascular and unspecified dementia (2.08%), renal failure (2.94%), intentional and unintentional injury, poisoning, and other adverse events (0.67%), all other conditions and causes (residual) (10.71%), and COVID-19 (26.03%). Therefore, the principal eigenvector of the Bayesian matrix B is

$s^T = (0.0031,\ 0.0100,\ 0.0256,\ 0.0630,\ 0.1439,\ 0.2327,\ 0.2648,\ 0.2569,\ 0.2904,\ 0.1839,\ 0.0269,\ 0.0156,\ 0.0386,\ 0.0124,\ 0.0079,\ 0.0208,\ 0.0294,\ 0.0067,\ 0.1071,\ 0.2603)$.

After identifying the principal eigenvector s, use the fmincon function from Matlab to find the minimum value of the constrained nonlinear multivariable function and solve the Bayesian matrix B. The first constraint condition, based on the characteristic of probability, is that all elements in the Bayesian matrix B must be in the interval [0, 1]. Therefore, all elements in the lb vector of the fmincon function are 0, and all elements in ub are 1. Additionally, the second constraint condition, based on the characteristic of collectively exhaustive events, is that the sum of each column in the matrix B is 1. Thus, the task is to construct the matrix Aeq and the vector beq. The third constraint condition, based on the properties of Markov matrices, is that $BS = SB$. Hence, it is vital to build another matrix Aeq and another vector beq. Lastly, based on the model from section 3, it is obvious that the $f(x)$ of the fmincon function is the maximization of the eigenvalue λ, which needs to be converted into the minimum problem by taking the negative value of λ.

The calculated Bayesian matrix B based on the above is shown below, and the specific value of each element in the matrix is presented in Figure 3. Figure 3 contains 192 values from left to right, and the Bayesian matrix B is filled column by column from left to right, top to bottom.

$$B = \begin{pmatrix} \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} & \begin{pmatrix} 0.0007 & \cdots & 0.0008 \\ \vdots & \ddots & \vdots \\ 0.2840 & \cdots & 0.2802 \end{pmatrix} \\ \begin{pmatrix} 0.0697 & \cdots & 0.3210 \\ \vdots & \ddots & \vdots \\ 0.0696 & \cdots & 0.2839 \end{pmatrix} & \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \end{pmatrix} \tag{10}$$

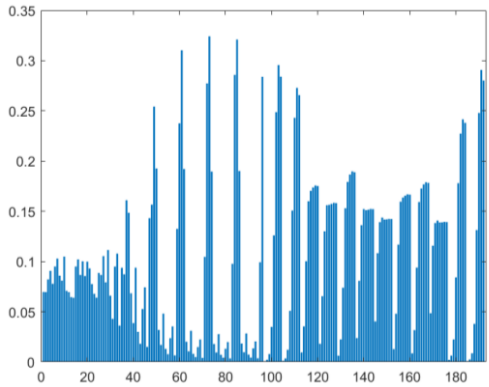

Figure 3. The value of 192 elements in B.

Based on the conditional probability given by B, $P(a_i|c_j)$ and $P(c_j|a_i)$ can be obtained. Therefore, the probability that a COVID-19 death patient had a health condition j in an age group i can be calculated through the following way:

$$P(a_i \cap c_j) = P(a_i|c_j) \times P(c_j) = P(c_j|a_i) \times P(a_i) \quad (11)$$

Sort the probabilities in descending order, the top 10 probabilities that a COVID-19 dead patients had a j health condition in an i age group are shown in Table 1:

Table 1. Hits for TOP 10 of the Results

COVID-19 dead patients had a j health condition in an i age group	Probability
85+ years old with respiratory diseases	0.0932
75-84 years old with respiratory diseases	0.0858
85+ years old with COVID only	0.0825
75-84 years old with COVID only	0.0757
65-74 years old with respiratory diseases	0.0722
65-74 years old with COVID only	0.0645
85+ years old with circulatory diseases	0.0552
75-84 years old with circulatory diseases	0.0502
65-74 years old with circulatory diseases	0.0447
55-64 years old with respiratory diseases	0.0366

5. Discussion and Conclusion

In this paper, according to the high-rank assumptions different from the previous low-rank assumptions, and using the characteristics of the Markov matrices, in the case of known principal eigenvectors, the Bayesian matrix B is completed, and 192 conditional probability values related to age groups and health conditions are obtained. Based on these conditional probabilities, it is easy to calculate the probability that a COVID-19 dead patients had a health condition j in an age group i. From the results, it should be noticed that there is a high probability that dead patients aged 85+ with respiratory diseases, followed by patients aged 75 to 84 with respiratory diseases. Besides, among the COVID-19 deaths, people aged 65+ who had circulatory diseases and those aged 55 to 74 who had respiratory diseases also accounted for a large proportion. Except for the top 10, patients over 65 with diabetes also have a high percentage of deaths. These findings can help hospitals to realize new medical decision-making and ensure the maximum overall benefit for all patients. In daily hospital management, doctors should give these patients a higher prioritization in the treatment of COVID-19.

Since COVID-19 is an RNA virus that mutates quickly, there are certain stability issues with this conclusion. In the future, experimental data could be limited to a period of time when the virus is relatively stable, and health conditions and age groups could be described in more detail to provide more stable and precise prioritization.

Acknowledgment

The work was supported by the National Natural Science Foundation of China (No.72271017).

References

[1] Haleem A, Javaid M, Vaishya R. Effects of covid-19 pandemic in Daily Life. Current Medicine Research and Practice. 2020;10(2):78–9.
[2] Who coronavirus (COVID-19) dashboard [Internet]. World Health Organization. World Health Organization; [cited 2023Feb24]. Available from: https://covid19.who.int/
[3] Choi W-Y. Mortality rate of patients with covid-19 based on underlying health conditions. Disaster Medicine and Public Health Preparedness. 2021;16(6):2480–5.
[4] John L. Hick DH. Crisis standards of care and covid-19: What did we learn? how do we ensure equity? what should we do? [Internet]. National Academy of Medicine. 2021 [cited 2023Feb24]. Available from: https://nam.edu/crisis-standards-of-care-and-covid-19-what-did-we-learn-how-do-we-ensure-equity-what-should-we-do/
[5] Hastie T, Friedman J, Tisbshirani R. The elements of Statistical Learning: Data Mining, Inference, and prediction. New York: Springer; 2017.
[6] Liu X, Wang X, Zou L, Xia J, Pang W. Spatial imputation for air pollutants data sets via low rank matrix completion algorithm. Environment International. 2020;139:105713.
[7] Li X, Yuan Y, Wang Q. Hyperspectral and multispectral image fusion via nonlocal low-rank tensor approximation and sparse representation. IEEE Transactions on Geoscience and Remote Sensing. 2021;59(1):550–62.
[8] Nie F, Hu Z, Li X. Matrix completion based on non-convex low-rank approximation. IEEE Transactions on Image Processing. 2019;28(5):2378–88.
[9] Hang Wu, Chihwen Cheng, Xiaoning Han, Yong Huo, Wenhui Ding, Wang MD. Post-surgical complication prediction in the presence of low-rank missing data. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2015;
[10] Gilks WR, Richardson S, Spiegelhalter DJ. Markov chain Monte Carlo in practice. Boca Raton etc.: Chapman and Hall; 1998.