WGU MSDA Program

Principal Component Analysis on Patient Data for Dimension Reduction

Write Up

Christopher Kamper

November 8th, 2023

**Part I: Research Question**

1. 1. Question: Can the medical data set be simplified to improve the efficiency of models without losing large amounts of variance?
   2. Goal: Use PCA to determine a more efficient way to describe the data by utilizing features calculated from the variables in the original data set. This will hopefully result in a data set with fewer features that can still represent over half of the original set.

**Part II: Method Justification**

1. 1. PCA will be used for this analysis as it is a dimensionality reduction method that will help minimize the amount of data needed to explain a data set. This process will project the original data points from a larger data set to a smaller sub-space. This is done by calculating principal components (PCs) from a standardized data set, that retains the essential information from the larger data set that explains the most variation in the data set. These PCs are found by calculating vectors from correlated variables so they can be represented in fewer, uncorrelated components. Since the correlated variables contribute to the same principal component, fewer are needed to describe the data set while retaining the variation. The variance described by each PC can be found by looking at the explained variance for each, which can help further reduce dimensionality as one can see how much variance is explained by each component and drop extras based on the threshold of variance that is desired. (Sharma, Aditya)
   2. One assumption of PCA is that the variables being worked with have a relation to each other. This means that while they may be independent, there is a correlation between the variables. (Great Learning Team)

**Part III: Data Preparation**

1. 1. Variables for PCA:
      1. Lat
      2. Lng
      3. Population
      4. Age
      5. Income
      6. VitD\_levels
      7. Doc\_visits
      8. Full\_meals\_eaten
      9. vitD\_supp
      10. Initial\_days
      11. TotalCharge
      12. Additional\_charges
      13. Item1: Timely Admission
      14. Item2: Timely Treatment
      15. Item3: Timely Visits
      16. Item4: Reliability
      17. Item5: Options
      18. Item6: Hours of Treatment
      19. Item7: Courteous Staff
      20. Item8: Evidence of Active Listening from Doc
   2. The data was standardized using the Standard Scaler from sklearn.preprocessing. Code is below:

#Set scaler.

scaler = StandardScaler()

#Fit scaler and transform the data frame.

medical\_std = scaler.fit\_transform(medical\_df)

#View standardized data.

print(medical\_std)

Cleaned data set is attached.

**Part IV: Analysis**

1. 1. The principal component matrix with all components is below:

Lat Lng Population Age Income VitD\_levels \  
PC1 0.008964 0.000729 0.009048 0.006664 0.001123 -0.003326   
PC2 -0.018479 -0.010546 0.029361 0.085471 -0.019874 -0.002735   
PC3 -0.001575 0.011236 -0.031260 0.693786 -0.016889 0.017117   
PC4 -0.016982 0.008909 -0.013232 -0.098148 0.022639 -0.027497   
PC5 -0.715350 0.272658 0.627549 0.010775 0.076125 -0.017846   
PC6 -0.108961 0.019969 0.050666 -0.022538 -0.372067 0.544875   
PC7 0.081546 -0.767536 0.401984 -0.003649 0.198204 0.244225   
PC8 -0.005286 0.369451 -0.210111 -0.006353 0.455009 0.210902   
PC9 0.031917 0.202410 -0.045156 -0.019084 -0.523500 0.376891   
PC10 -0.066441 0.095028 -0.105870 -0.013638 0.493306 0.157951   
PC11 -0.043207 -0.023595 0.054997 -0.017500 -0.283065 -0.652159   
PC12 0.085424 0.029438 0.097334 -0.005064 -0.077394 0.050408   
PC13 -0.659295 -0.380145 -0.593603 0.004115 -0.066011 0.011787   
PC14 0.137383 0.055536 0.127093 0.015725 0.000290 0.002867   
PC15 -0.022570 0.011899 -0.016380 -0.001241 -0.006717 -0.040734   
PC16 0.016120 0.006496 -0.015506 0.008291 0.017745 -0.014759   
PC17 -0.004123 -0.005812 -0.026803 -0.019302 0.001229 -0.021294   
PC18 -0.002693 0.008716 -0.004870 0.000438 0.003807 0.016278   
PC19 0.008663 -0.004990 0.016334 0.706200 0.002168 -0.002585   
PC20 0.001371 -0.000410 -0.000628 0.026317 0.001269 -0.001504

Doc\_visits Full\_meals\_eaten vitD\_supp Initial\_days TotalCharge \  
PC1 0.007237 -0.000445 -0.004983 -0.026359 -0.024766   
PC2 -0.006277 -0.020911 0.025133 0.699727 0.700950   
PC3 0.015897 0.028769 0.014015 -0.089468 -0.079014   
PC4 0.003542 -0.028684 -0.006175 0.010254 0.008759   
PC5 0.016329 -0.103207 0.029229 -0.023838 -0.021789   
PC6 -0.095667 0.596242 -0.417751 0.010095 0.013175   
PC7 0.347137 0.048820 0.132099 -0.011412 -0.009877   
PC8 0.719183 0.078602 -0.196382 0.022615 0.022958   
PC9 0.304106 -0.183196 0.638107 -0.014473 -0.012667   
PC10 -0.334494 0.507546 0.579800 0.006144 0.006870   
PC11 0.377870 0.567729 0.147593 0.000711 0.002385   
PC12 -0.026542 0.060375 0.034318 -0.003515 -0.004577   
PC13 0.060301 -0.059434 0.015021 0.002049 0.001058   
PC14 0.017318 0.045221 -0.029838 0.011071 0.009115   
PC15 -0.023515 -0.011061 0.012995 0.001611 -0.000080   
PC16 0.007367 0.001091 -0.003623 0.002777 0.003582   
PC17 -0.009520 0.015606 0.010645 -0.000555 -0.000392   
PC18 -0.001964 0.001181 0.000072 -0.005311 -0.006814   
PC19 0.000377 0.010994 0.000815 0.031452 -0.031619   
PC20 -0.001119 -0.001604 -0.000594 -0.706294 0.706460

Additional\_charges Item1: Timely Admission Item2: Timely Treatment \  
PC1 0.006184 0.454577 0.428160   
PC2 0.085845 0.006674 0.017249   
PC3 0.694332 -0.043366 -0.044699   
PC4 -0.093751 -0.293643 -0.289666   
PC5 0.013358 0.008218 0.014722   
PC6 -0.017208 0.005306 -0.001710   
PC7 0.013838 -0.009111 0.000667   
PC8 -0.010435 0.002323 -0.009784   
PC9 -0.026959 0.010957 0.011136   
PC10 -0.009298 0.006548 -0.008883   
PC11 -0.006393 -0.015918 -0.007929   
PC12 0.011314 -0.097865 -0.149272   
PC13 -0.018771 0.010490 0.005771   
PC14 0.003772 0.071614 0.130466   
PC15 0.008660 -0.012344 -0.063825   
PC16 0.002611 0.084337 0.097881   
PC17 0.009449 0.183281 0.626198   
PC18 0.002524 -0.805563 0.533580   
PC19 -0.705729 -0.001065 0.008599   
PC20 -0.036762 -0.001476 0.000330

Item3: Timely Visits Item4: Reliability Item5: Options \  
PC1 0.395023 0.152292 -0.190140   
PC2 0.011804 -0.000557 -0.009686   
PC3 -0.046813 0.071517 -0.084670   
PC4 -0.291901 0.552995 -0.577254   
PC5 -0.012651 0.007619 0.014108   
PC6 -0.018678 0.032565 -0.023798   
PC7 -0.027553 -0.003925 -0.013354   
PC8 -0.017295 -0.053598 -0.001028   
PC9 0.000482 0.040362 -0.012547   
PC10 0.019819 0.019042 -0.005661   
PC11 0.017264 0.004152 0.003406   
PC12 -0.217335 -0.370355 0.122270   
PC13 0.037138 0.024729 0.025606   
PC14 0.200596 0.351476 -0.060839   
PC15 -0.234036 -0.390109 -0.133386   
PC16 -0.429027 0.483268 0.695921   
PC17 -0.623507 -0.112001 -0.303276   
PC18 0.193324 -0.011018 0.093439   
PC19 -0.018361 -0.012782 -0.010145   
PC20 -0.000793 -0.001332 -0.000549

Item6: Hours of Treatment Item7: Courteous Staff \  
PC1 0.410212 0.356415   
PC2 0.012976 0.026538   
PC3 0.010672 0.028157   
PC4 0.163297 0.170836   
PC5 -0.010708 -0.005799   
PC6 0.035566 0.022246   
PC7 0.030966 0.009989   
PC8 0.006211 0.002593   
PC9 0.011664 -0.011634   
PC10 0.003249 -0.009638   
PC11 -0.005292 0.026642   
PC12 -0.040346 0.061297   
PC13 -0.024775 -0.150203   
PC14 -0.053857 -0.832830   
PC15 0.796013 -0.331065   
PC16 0.272067 0.068341   
PC17 -0.271231 -0.061673   
PC18 0.125938 0.051523   
PC19 0.000284 0.007363   
PC20 0.000366 -0.001772

Item8: Evidence of Active Listening from Doc   
PC1 0.312572   
PC2 0.010127   
PC3 0.009723   
PC4 0.168805   
PC5 0.006366   
PC6 -0.076291   
PC7 -0.029184   
PC8 0.036641   
PC9 -0.053742   
PC10 -0.004491   
PC11 -0.022251   
PC12 0.855293   
PC13 0.173856   
PC14 0.268416   
PC15 -0.155684   
PC16 0.038175   
PC17 -0.038733   
PC18 0.033865   
PC19 0.002590

PC20 0.001108

1. Using the elbow rule, the scree plot shows that the drop off of variance explained occurs after 6 components. (Sanchita Mangale)

A graph showing the growth of a medical scree plot

Description automatically generated

1. Variance of each component:
   1. Variance described by PC1 = 14.77 %
   2. Variance described by PC2 = 9.97 %
   3. Variance described by PC3 = 8.58 %
   4. Variance described by PC4 = 8.25 %
   5. Variance described by PC5 = 6.14 %
   6. Variance described by PC5 = 5.2 %
2. Total variance of components:

The total variance described is 52.91 %.

1. The result of this analysis shows that 6 principal components can describe over 50% of the variance in this data set. This shows the analysis has been successful in identifying a smaller set of components that can explain the majority of the variance in the data. Since these 6 components represent 52.91% of the variance, one could run further analysis such as predictive modeling in a more efficient manner with confidence in knowing they are using much less data while retaining most of the information.

**Part V: Attachments**

1. Web Sources for Code:

Sharma, Aditya. “Principal Component Analysis (PCA) in Python Tutorial.” Datacamp.com, DataCamp, 2020, [www.datacamp.com/tutorial/principal-component-analysis-in-python](http://www.datacamp.com/tutorial/principal-component-analysis-in-python). Accessed 27 Oct. 2023.

“PCA with Data Mining II - D212.” *Panopto*, 2015, wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=a7592d75-cc72-47fd-8a9a-b07a00efff97. Accessed 31 Oct. 2023.

1. Sources:

Sharma, Aditya. “Principal Component Analysis (PCA) in Python Tutorial.” Datacamp.com, DataCamp, 2020, [www.datacamp.com/tutorial/principal-component-analysis-in-python](http://www.datacamp.com/tutorial/principal-component-analysis-in-python). Accessed 27 Oct. 2023.

SANCHITA MANGALE. “Scree Plot - SANCHITA MANGALE - Medium.” *Medium*, Medium, 28 Aug. 2020, sanchitamangale12.medium.com/scree-plot-733ed72c8608. Accessed 27 Oct. 2023.

Great Learning Team. “Understanding Principal Component Analysis and Their Applications.” *Great Learning Blog: Free Resources What Matters to Shape Your Career!*, 18 May 2020, [www.mygreatlearning.com/blog/understanding-principal-component-analysis/#assumptions](http://www.mygreatlearning.com/blog/understanding-principal-component-analysis/). Accessed 27 Oct. 2023.