Using Contrastive Principal Component Analysis to Identify Post-acute Sequelae of SARS-CoV-2 Infection Subphenotypes

Xiaokang Liu [1,2], PhD, Yishan Shen [2*], MS, Naimin Jing [3] , PhD,
Christopher B. Forrest[4], MD, PhD, Yong Chen [2**], PhD
[1] Department of Statistics, University of Missouri, Columbia, MO 65211, USA
[2] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA
[3] Biostatistics and Research Decision Sciences, Merck & Co., Inc, Kenilworth, NJ 07033, USA
[4]Applied Clinical Research Center, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[*] co-first author
[**] corresponding author

Background

The post-acute sequelae of SARS-CoV-2 infection (PASC), commonly known as "long COVID", refers to a range of persistent or new symptoms that emerge after the acute phase of COVID-19 infection[1]. These symptoms can endure for weeks or months following the initial infection and can affect various body systems. Moreover, the PASC symptoms can vary widely between individuals, which brings challenges to the diagnosis and treatment of PASC patients. To gain more understanding of dominant symptom co-occurrence patterns of PASC and develop effective treatments, identifying subtypes (also known as subphenotypes) of PASC is of great interest to both health care providers and patients. Traditional clustering methods to find subphenotypes of a complex syndrome disease include latent class analysis[2] and agglomerative hierarchical clustering [3]. However, since the highly heterogeneous spectrum of PASC clinical features can overlap with features of other diseases, the identified subphenotypes may not be specific to PASC. Hence, with electronic health records (EHR) for both COVID-19 test-positive and test-negative patients extracted from the PEDSnet[4,5] COVID-19 Database, we applied a contrastive principal component analysis method (cPCA) [6] to help derive PASC subphenotypes for children[7]. This study aims to provide more insights into PASC and facilitate tailored interventions for affected children.

Methods

cPCA is a method to identify prominent trends that are specific to a target dataset $\{x_i\}$, which is of the main interest to the researchers, relative to a comparison background dataset $\{y_i\}$. Specifically, the method takes both datasets as inputs and calculates their respective variance-covariance matrices Σ_x and Σ_y. Then, the contrastive projection directions are the vectors v that maximize $v^{\mathrm{T}}(\Sigma_x - \lambda\Sigma_y)v/v^{\mathrm{T}}v$ where λ determines the desired contrast level. Therefore, the method produces subspaces that capture a significant amount of variation within the target data $\{x_i\}$, while exhibiting minimal variation in the background $\{y_i\}$. The features within this subspace encapsulate structures specific to $\{x_i\}$. We project the target data onto this subspace and use k-means to discover the additional clustering patterns unique to the target data relative to the background. In our PASC analysis, EHR of COVID-19 test-positive patients form the target dataset, while the EHR of COVID-19 test-negative patients constitute the background dataset, which contains information regarding general disease patterns not specific to PASC.

The authors have released a python package for cPCA with documentation and examples at https://github.com/abidlabs/contrastive. The input includes a target data and a background data that have the same format but are from different cohorts. The returned data is the projected target data. The user can specify the number of top contrastive principal components used in the analysis. The parameter λ that controls the degree of contrast can either be provided by the user or decided by the user after inspecting results returned by experimenting with multiple λ's. The built-in plotting and interactive GUI help the user see how the target data points move as the user change the value of the contrast parameter, which provides insights about the possible groups in the target data.

Results

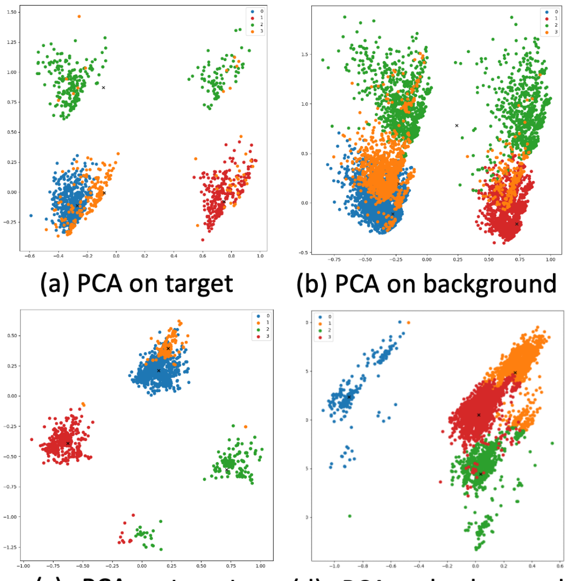
The cPCA method proved effective in eliminating background information and identifying projection directions specific to PASC, leading to distinct clustering results compared to those obtained by solely applying PCA on the target data. Figure 1 visually represents the clustering results obtained using k-means based on principal components derived from various methods, including: (a) PCA applied solely on the target data, (b) PCA applied solely on the background data, (c) cPCA applied on both datasets, and (d) projecting the background data onto the directions obtained from applying cPCA on both datasets.

The observations can be summarized as follows:

I. The principal variational directions in the target data and the background data are similar, resulting in similar clustering results between both datasets when PCA is applied, as depicted in panels (a) and (b) of Figure 1.

II. cPCA can find PASC-specific projection directions which lead to well-separated clusters in the target data, and these directions cannot well separate clusters in the background data, as evidenced by panel (c) and (d) in Figure 1.

Figure 2 further reinforces these findings by illustrating the distances between clusters obtained using different methods on different datasets through the utilization of multi-dimensional scaling [8]. The figure indicates that while the four clusters identified in the target data based on applying PCA solely to the target data align well with the four clusters in the background data derived from applying PCA solely to the background data, representing general disease patterns shared by both test-positive and test-negative patients, cPCA assists in discovering unique clusters that are specific to PASC.


(a) PCA on target (b) PCA on background

(c) cPCA on target (d) cPCA on background

Figure 1. Clusters found in each dataset by applying k-means to principal components derived by different methods: Panel (a) apply PCA on target data alone; Panel (b): apply PCA on the background data alone; Panel (c): apply cPCA on both datasets; Panel (d): project the background data to directions obtained by applying cPCA on both datasets.

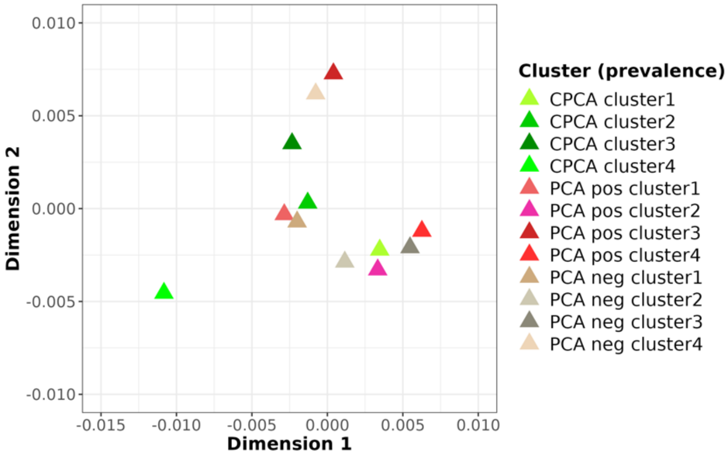


Figure 2. The distance between clusters found by different methods on each dataset. cPCA clusters: clusters in the target data found by k-means using principal components learned by cPCA. PCA pos clusters: clusters in the target data found by k-means using principal components learned by applying PCA to the target data alone. PCA neg clusters: clusters in the background data found by k-means using principal components learned by applying PCA to the background data alone.

The application of cPCA has successfully identified four distinct clusters, each representing different disease patterns within PASC. Figure 3 illustrates the prevalence of various symptoms within each cluster, providing further insights into their characteristics. Specifically, class 1 corresponds to a mild disease presentation, most patients in class 2 develop both anxiety disorder and teeth and gingiva disorders, class 3 comprises patients exhibiting COVID-19-related symptoms, and class 4 represents the upper respiratory infection subpopulation.

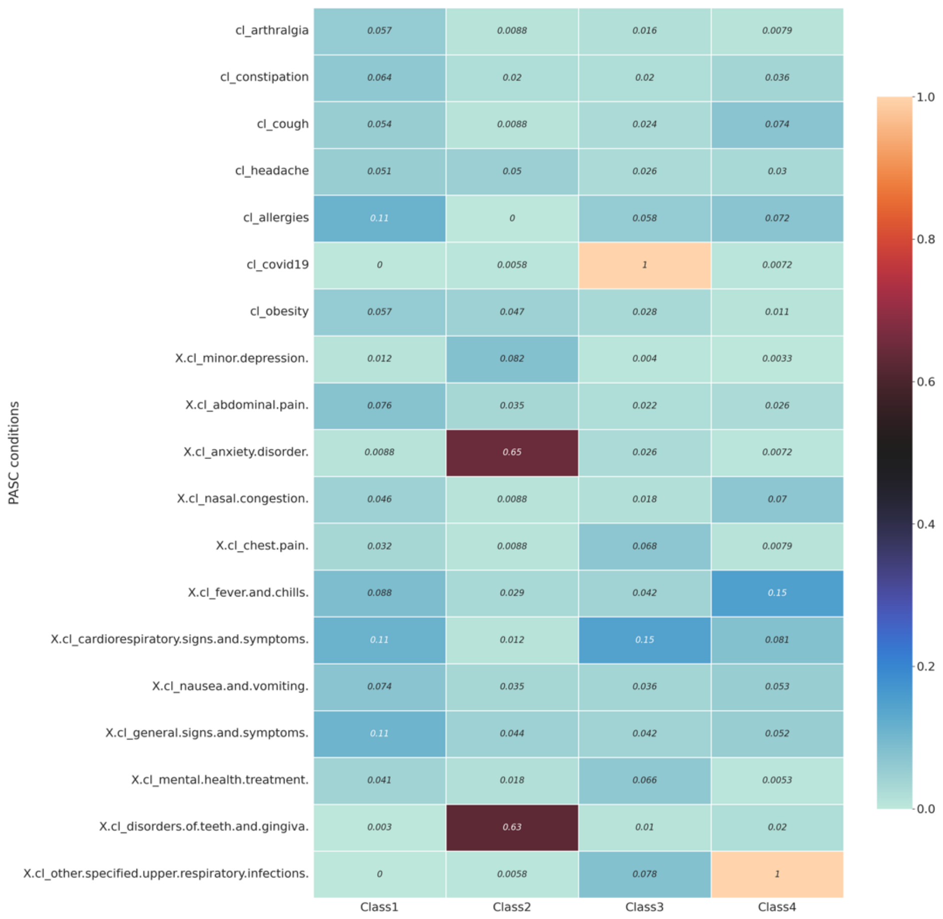


Figure 3. Subphenotypes identified using cPCA. Each column represents a subphenotype, and it is characterized by the prevalence of several symptom indicator variables (rows). The cells are color-coded to represent the prevalence of the variables, with the legend to the right showing the scale of the colors used. The numbers in cells are the corresponding prevalence values.

Conclusion

cPCA is a powerful tool to identify target data-specific signals from a target dataset by effectively eliminating background information contained in a background dataset. Our analysis results indicate that the general disease patterns shared by both test-positive and test-negative patients are the dominant disease trends in the target dataset, which may fail the subphenotyping task if solely using the test-positive population to learn PASC subphenotypes. In contrast, cPCA can help us find unique PASC-specific subphenotypes.

Our study is a successful practice of generating reproducible and reliable evidence using real-world data in addressing important public health and biomedical question, and our findings underscore the importance of utilizing advanced techniques to achieve accurate and informative subphenotyping of PASC, ensuring a more nuanced understanding of the condition and paving the way for tailored interventions and improved patient care.

This method can be directly applied to any datasets that are harmonized by OMOP CDM with no further pre-processing or variable selection needed. Further evaluation of the method on OHDSI studies for different diseases and against different datasets will be conducted in the future.

References

1. Rao S, Lee GM, Razzaghi H, Lorman V, Mejias A, Pajor NM, Thacker D, Webb R, Dickinson K, Bailey LC, Jhaveri R, Christakis DA, Bennett TD, Chen Y, Forrest CB. Clinical features and burden of post-acute sequelae of SARS-CoV-2 infection in children and adolescents. JAMA Pediatr. 2022;176(10):1000–1009.
2. Bandeen-roche K MD, Zeger SL, Rathouz PJ. Latent variable regression for multiple discrete outcomes. Journal of the American Statistical Association. 1997;92(440):1375-1386.
3. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?. Journal of classification. 2014;31:274-95.
4. Forrest CB, Margolis PA, Bailey LC, et al. PEDSnet: a National Pediatric Learning Health System. J Am Med Inform Assoc. 2014;21(4):602-6.
5. Forrest CB, Margolis P, Seid M, Colletti RB. PEDSnet: how a prototype pediatric learning health system is being expanded into a national network. Health Aff (Millwood). 2014;33(7):1171-7.
6. Abid A, Zhang MJ, Bagaria VK, Zou J. Exploring patterns enriched in a dataset with contrastive principal component analysis. Nature communications. 2018;9(1):2134.
7. A part of the NIH Researching COVID to Enhance Recovery (RECOVER) initiative https://recovercovid.org/.
8. Chen CH, Härdle W, Unwin A, Cox MA, Cox TF. Multidimensional scaling. Handbook of data visualization. 2008:315-47.