Applying PCA to Improve Accuracy of VDH Health Opportunity Indicies

Student: Chelsea Qian Advisor: Prof. Aaron Schroeder

Background and Model Overview

- The model aims to measure Health Opportunity Indicies (HOIs) by integrating multiple socioeconomic indicators.
- Based on linear model, we assign weights to each measures to calculate annual HOI scores for each county.
- Interactive visualizations are used to display these indicies, enabling comparisons across regions and providing summary statistics on each indicies.

Goals in this project

- Improve model accuracy by refining Health Opportunity Indicies through PCA analysis and optimization of raw county data, specifically for year 2019 and 2020.
- Aims to offer insights into regional disparities in health opportunities and supports data–drivien policy–making to enhance health equity.

Method Used

- Principal Components Analysis (PCA): a linear dimensionality reduction and machine learning method used to simplify a large dataset into smaller set while still maintaining significant patterns and trends.

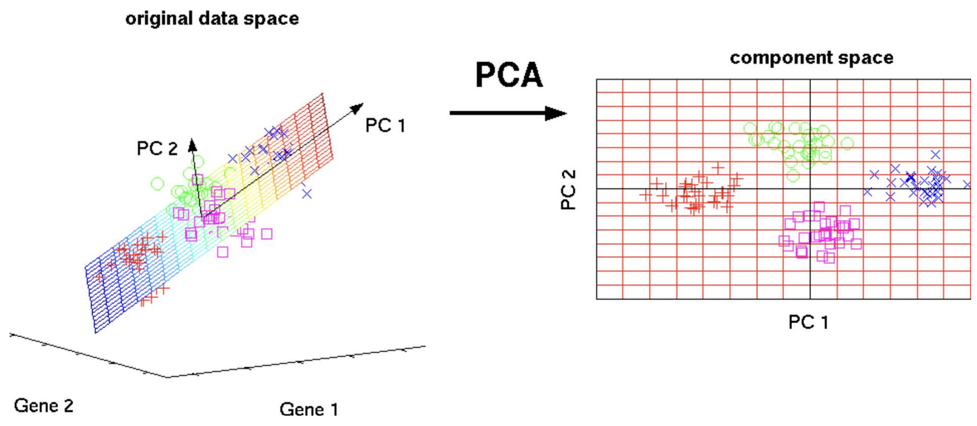

Figure 2: PCA visual demonstration

Analysis Process

- 1. Standardize the data: adjust the dataset to have a mean of zero and standard deviation of one to ensure all variables are on the same scale.
- 2. Compute the covariance matrix to identify correlations: understand how features vary with respect to each other

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14
**Accees to Care	0.023	0.391	0.594	0.150	0.486	0.326	-0.095	0.262	-0.091	0.138	0.120	-0.057	-0.074	-0.009
Zscore_Education**	0.527	0.665	0.155	0.204	0.105	-0.189	-0.095	-0.028	0.116	-0.011	-0.080	0.035	0.365	-0.051
Zscore_Employment Access	0.870	0.029	0.108	0.087	0.136	-0.062	0.277	-0.026	0.050	-0.115	-0.090	-0.026	-0.054	0.308
Zscore_Income Inequality	-0.137	-0.436	0.512	0.498	-0.212	0.096	0.044	-0.247	0.346	0.157	0.107	0.062	-0.062	-0.014
Zscore_Labor Force Participation	0.639	0.221	-0.363	-0.271	0.016	0.013	-0.182	0.111	0.364	0.064	0.391	0.011	-0.044	0.019
Zscore_Population Density_RC	-0.737	0.125	-0.083	0.100	-0.080	0.171	-0.304	0.232	0.363	-0.127	-0.263	-0.127	-0.001	0.099
**Spatial Segregation	-0.082	-0.393	-0.430	0.287	0.601	-0.222	-0.292	-0.234	-0.026	0.116	-0.058	-0.029	-0.042	0.022
Zscore_Walkability	0.354	-0.206	-0.293	0.622	-0.257	-0.122	-0.034	0.490	-0.173	0.061	0.055	0.013	-0.010	0.001
Zscore_Affordability_RC*	0.595	-0.476	-0.074	-0.207	-0.079	0.336	-0.072	0.035	0.037	0.389	-0.256	0.145	0.070	0.019
Zscore_Environmental_RC*	-0.841	0.142	0.010	-0.059	-0.172	-0.029	-0.132	-0.074	-0.193	0.230	0.218	0.005	0.156	0.230
Zscore_Food Access_RC*	0.358	0.417	-0.240	0.387	-0.196	0.496	-0.171	-0.373	-0.123	-0.138	0.052	0.002	-0.028	0.005
Zscore_Townsend_RC*	-0.374	0.795	-0.111	0.059	0.051	-0.112	0.027	0.036	0.035	0.038	-0.097	0.400	-0.151	0.023
Zscore_Mobility_RC*	-0.490	0.258	-0.435	0.195	0.166	0.198	0.575	0.002	0.154	0.170	0.023	-0.099	0.069	-0.035
ZScore_Incarceration_RC*	0.296	0.762	0.012	-0.054	-0.238	-0.240	-0.061	-0.107	-0.046	0.291	-0.148	-0.255	-0.161	-0.026

Figure 3&4: Example of covariance matrix and calculated Covariance matrix

- 3. Calculate the Eigenvalues and Eigenvectors: derived from the covariance matrix to identify the principal components.

2019	F1	F2	F3	F4	F5
Eigenvalue	7.644	2.812	1.438	1.189	0.932
Variability (%)	44.964	16.541	8.459	6.993	5.484
Cumulative %	44.964	61.505	69.964	76.957	82.441

2019	Weights	eight Adjusted
F1	44.964	0.58427534
F2	16.541	0.21494022
F3	8.459	0.1099181
F4	6.993	0.09086634
	76.957	1

2020	F1	F2	F3	F4	F5
Eigenvalue	6.687	3.178	1.745	1.177	0.909
Variability (%)	39.336	18.695	10.263	6.923	5.348
Cumulative %	39.336	58.032	68.295	75.218	80.566

2020	Weights	eight Adjusted
F1	39.336	0.5229646
F2	18.695	0.24854621
F3	10.263	0.13644492
F4	6.923	0.09204427
	75.218	1

Figure 4&5: Eigenvalues and Weights for 2019 and 2020

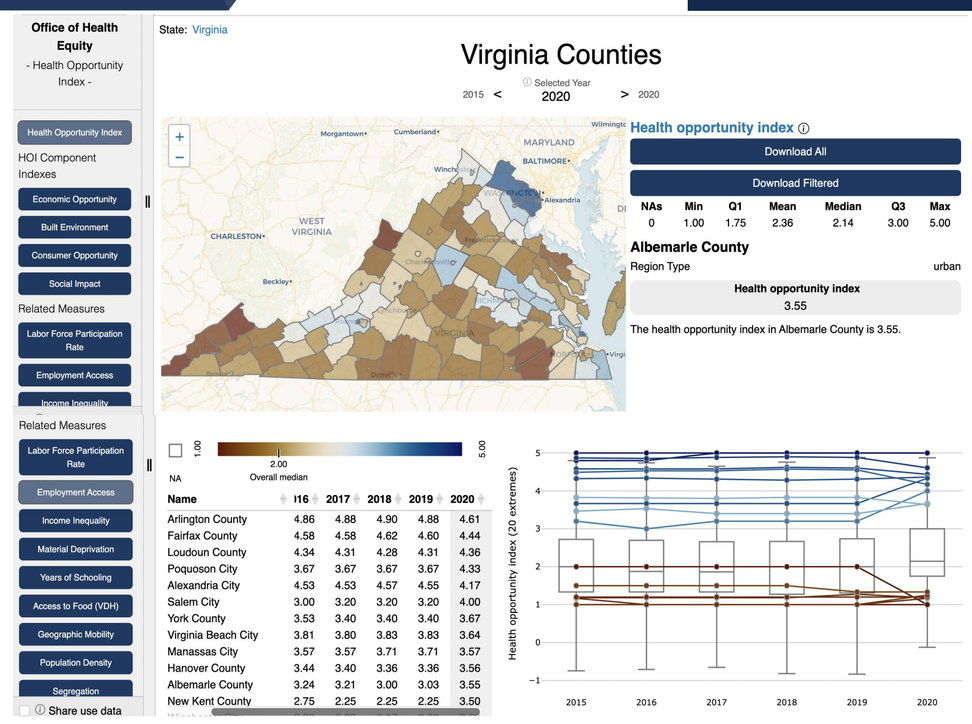

Figure 1: Model Visualization

Analysis Process Continued

- 4. Select the principal components: Choose the Top principal components based on Eigenvalues.
- 5. Transform the data & Rotation: Project the original data onto the new feature space defined by principal components. Orthomax (varimax) rotation was applied to align the data more closely with the principal components

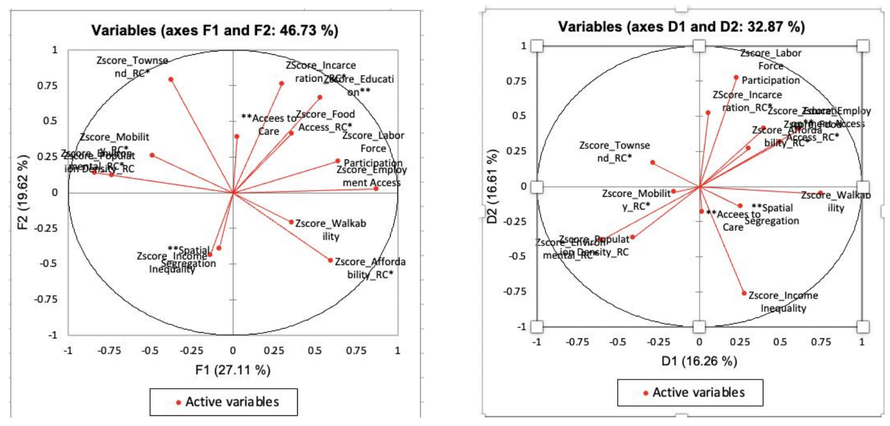

Figure 7&8 Factors before and after rotation[a]

- 6. Obtain the final indicies: aggregate the transformed data for each related measure for each county to derive the final Health Opportunity Index

Results

- Updated statistics for each county for year 2019&2020

	D1	D2	D3	D4
51091970100	-1.207	-0.836	0.229	0.244
51017920101	-0.953	-0.764	0.253	0.992
51017920102	-0.699	-0.909	0.217	-0.101
51015070100	-0.869	-1.640	-0.436	-0.046
51015070802	-0.657	0.019	-0.314	0.798
51015070801	-0.532	-0.556	0.084	1.034
51015070200	-0.923	0.094	0.018	1.374
51165011200	-0.532	0.070	-1.163	1.400
51165011000	-1.543	0.048	-0.830	0.250
51005080301	-1.546	-0.130	0.718	0.865
51171040201	-0.674	-0.402	0.187	0.527
51005080302	-1.301	-0.908	0.423	0.543

Figure 9: Example of result statistics

Future Discussions

- Incorporate additional indicators: integrate more socioeconomic, environmental, and healthcare access indicators to enhance the robustness of HOIs.
- Temporal analysis: Conduct longitudinal analysis for trends and changes over time, identify factors contributing to improvements or declines.

References

- *Principal Component Analysis (PCA) explained. Built In. (n.d.). https://builtin.com/data-science/step-step-explanation-principal-component-analysis*
- *Exploratory factor analysis: Rotation. (n.d.). https://www.ibm.com/docs/en/spss-statistics/beta?topic=analysis-exploratory-factor-rotation*

UNIVERSITY of VIRGINIA
BIOCOMPLEXITY INSTITUTE