**Summary of new method of identifying key industries:**

**a principal component analysis**

In this article discusses the application of dimensionality reduction, in particular P.C.A. (Principal Component Analysis), to three benchmark input-output tables of the U.S.A.(for the years 2002, 2007, 2012) as well as for the year 2019 using data given by the U.S. Bureau of Economic Analysis (BEA). The dimensions of each were reduces to 70x70 matrices eliminating the variables “housing industry” and “fictitious household industry”. Dimensionality reduction is a way to reduce the complexity of a model, its origins can be traced back to the physiocrats (tableau economique) and similar techniques (the Schur, SVD ) have been identified in plethora of other works since then.

The article describes/applies P.C.A. as a dimensionality reduction technique that transforms data (points in space) to lower dimension through linear combinations of the original data. Basically, it transforms correlated variables into fewer uncorrelated variables and then projects the original data into the reduced P.C.A. space using the eigenvectors/eigenvalues of the variance/covariance matrix (Principal Components). The resulting data captures most of the variance of the original data set.

Even though P.C.A. has not been used extensively in input-output analysis, it holds a key advantage in the sense that it is able to identify the relative importance of the industries operating in the economy.

The author goes on to describe how a P.C.A. is applied to input-output resulting in the computation of the Principal Components and their transformation into the PCA space using the variance/covariance matrix.

* PC: represents a new set of variables that are uncorrelated and account for the maximum variance in the data, It is derived from the eigenvectors of the variance/covariance matrix H' \* H
* : variance/covariance matrix (on its diagonal are the eigenvalues ranked max to min )
* (70-1): degrees of freedom

From the resulting we keep the eigenvectors corresponding to the top two eigenvalues, hence reducing the dimensionality to 70x2 matrix.

The text goes on to discuss the application of P.C.A. in the context of identifying Key industries, it then goes on to compare P.C.A with the Leontief inverse as well as the estimates for forward and backwards linkage (FL, BL respectively). Computing the Pearson correlation efficient a strong and positive relationship between the first PC and the unweight BL of each industry is revealed. Still FL remains important for understanding industry interrelationships and economic structural changes. Using both the first and second PCs (to account for FL) is essential for a comprehensive analysis.

Next the 70 industries are clustered based on similarity/homogeneity. First the k-means criterion (divides n observation into k clusters) is used to part our data then we choose the n of clusters based on the highest Silhouette Score (how well the data fits in its assigned cluster) which in our case is three for every year. K-means is based on the principal of minimizing the Euclidean distance (similarity measure) withing in each cluster, meaning from each point to the “center” of the cluster. Then the industries are classified into the three clusters based on their Euclidean distance to each centroid.

To test the effectiveness of P.C.A. the author goes on to compare its findings with those of total BL and FL. Whilst neither of them taken individually produce anything like the results in ranking as the P.C.A., both of them taken together produce a similar ranking, this supports our results as meaningful.

In effect it is concluded that the P.C.A. stands out as a valuable tool for analyzing input-output data, refining industry rankings and revealing cluster structures since it is undoubtedly better at capturing the variance associated with each of the industries, (because it allows for said variance to be measured as the distance of a “data point “ from zero) than traditional method. As well as giving us the analytical advantage by organizing the data into well-defined clusters and dendrograms(number of branches based on the k-mean criteria) and yet having similar raking as the BL,FL approach. The author encourages for further research on various levels.