PCA and Isomap Analysis **Aviva Prins**
Gang Reduction and Youth Development Spring 2018

Introduction

The Gang Reduction and Youth Development (GRYD) program in Los Angeles has a program for adolescents at risk of joining gangs. GRYD uses uses a questionnaire ('Youth Services Eligibility Tool', or YSET) to evaluate the eligibility of students for this program. Students are sorted into three categories: low risk, at risk, and high risk. The program aims to provide resources for the at risk group.

The dataset $\mathcal{D}$ contains the responses of $N \sim 10,000$ students. The questionnaire has 100 questions, of which only 56 are used for risk assessment. To calculate risk factor, a student's score is calculated using nine thematic scales. A student scoring in less than three of the scales is considered to have low risk factor, and a student scoring in more than five of the scales is considered to have high risk factor.

There may be other methods that are more effective at classifying students into three groups. I investigate visualizations of the data in order to identify features of our dataset and to be able to compare different methods. If each response can be thought of as a point in real coordinate space, the problem becomes whether I can project the space onto two dimensions without loss of information. Any useful result will plot at risk responses between low risk and high risk. A great result will cluster the three groups. In this case, the visualization can be considered as a classification technique.

Methods

Instead of doing analysis directly, I use the graph Laplacian L on either the scales or the questions. Considering that there are nine scales and 56 questions, a response $x_i \in \mathbb{R}^9$ or $x_i \in \mathbb{R}^{56}$, respectively. The dataset can be written as a graph with nodes x_i and weighted edges w_{ij} corresponding to the similarity in answering the questionnaire. The weights w_{ij} correspond to a Gaussian distribution, scaled according to the Euclidean distance between responses x_i and x_j and a parameter σ. Then the corresponding graph Laplacian $L = D - W$, where:

$$W_{ij} = e^{\frac{-\|x_i - x_j\|^2}{2\sigma^2}}$$

$$D = \text{diag}\left(\sum_{j=1}^{m} W_{ij}\right)$$

The graph Laplacian is positive semidefinite, meaning we can decompose it into eigenvalues $\lambda_i \geq 0$ such that $Lv_i = \lambda_i v_i$.

I primarily focus on two methods to reduce the dimensionality of the dataset $\mathcal{D} = \{x_i\}_{i=1}^N$ onto $\mathbb{R}^2$: Principle Component Analysis (PCA) and Isomap. PCA linearly projects the data onto the directions in which the sample variance is maximized. Isomap is an extension of Multidimensional Scaling (MDS): it uses only the nearest k neighbors to compute the pair-wise distances between all points in the dataset, then uses MDS to nonlinearly project them.

PCA works as follows. Let $\mathbf{V} \subset \mathbb{R}^2$ be the subspace spanned by the basis of the first two unit eigenvectors corresponding to the largest eigenvalues of $L_{d\times d}$, λ_1 and λ_2. Then we can project a point x onto $\mathbf{V}$:

$$\hat{x} = \text{proj}_{\mathbf{V}}(x) = \bar{x} + \sum_{i=1}^{2} \langle x - \bar{x}, u_i \rangle u_i$$

where $\bar{x}$ is the sample mean. The error using least squares is

$$\frac{1}{N} \sum_{n=1}^{N} \|\hat{x_n} - \bar{x}\|^2 = \sum_{j=3}^{d} \lambda_j$$

so the projection minimizes the error by taking the largest two eigenvalues.

Isomap is a nonlinear dimensionality reduction method that extends metric multidimensional distances (MDS). The algorithm has three steps. First, it performs k-nearest neighbors to identify the contour of the manifold. Second, it estimates the geodesic distances by calculating the shortest paths between points. Finally, Isomap performs MDS on the new distance matrix.

Results

Though Principle Component Analysis (PCA) was investigated before this quarter, the majority of my time was spent debugging this code. Thus, even though the results are not new, I will be discussing them in depth.

Figure 1 shows the results of doing PCA and Isomap on the scales. The responses are clearly clustered into three groups, however these three groups do not correspond to Risk Factor (RF). Previous investigation on PCA has found that these clusters relate to one of the scales which has only two questions. I suspect that the underlying factor for clustering the Isomap method is the same.

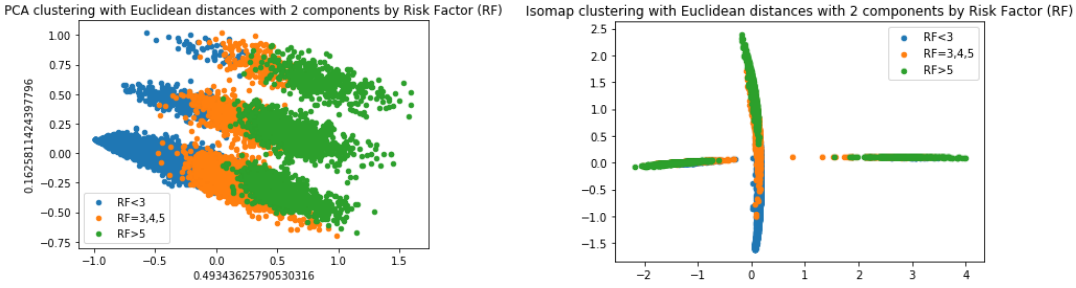

Figure 1: PCA (left) and Isomap (right) on the scales

PCA on the questions yields one large cluster with a clear progression from low Risk Factor to high Risk Factor, as seen in Figure 2. Unfortunately, it appears that there is a lot of overlap which is revealed in Figure 3.

Isomap on the questions has the same qualities: there is one large cluster with a trend corresponding to risk factor (see Figure 4) but there is a large amount of overlap, as seen in Figure 5. Unless indicated otherwise, I used the default $k = 5$ for k-nearest neighbors. I do not know what values of k give the best results; I tried $k = 50$ but the plot does not cluster the points well.

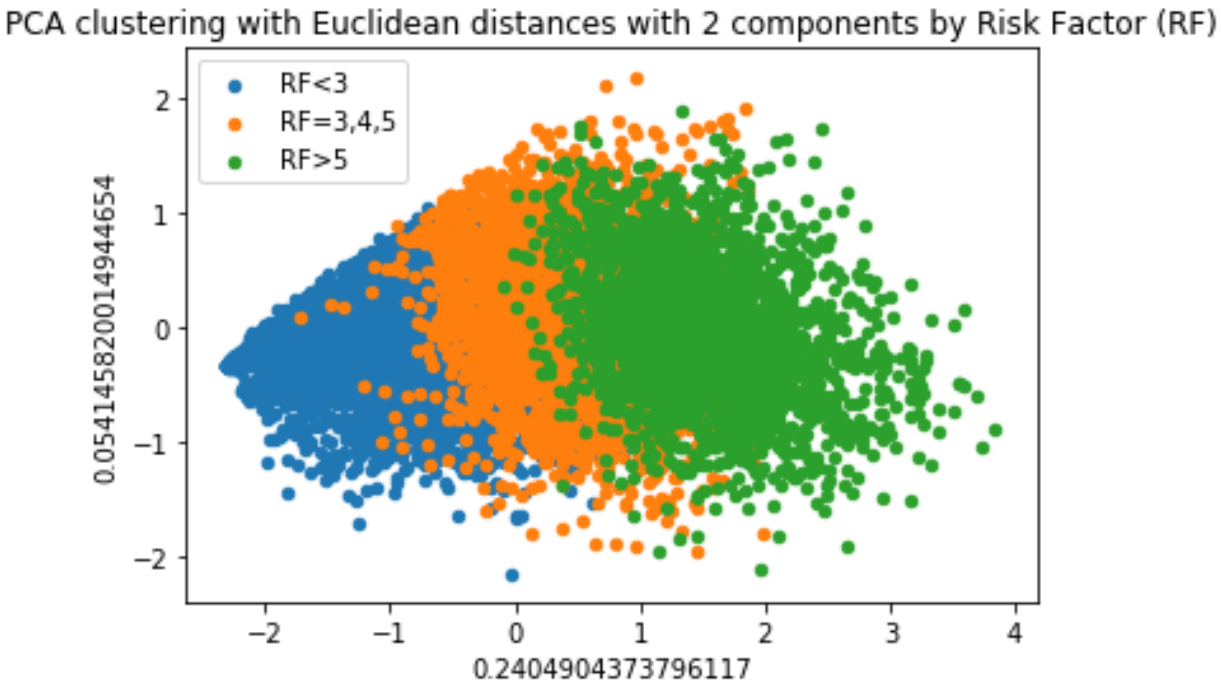

Figure 2: PCA on the questions

Since other methods are implemented in the same python module, I tried Locally Linear Embedding, Spectral Embedding, and t-distributed Stochastic Neighbor Embedding in Figure 6. None of these results are promising, so I did not investigate them further.

In conclusion, both PCA and Isomap project in some way that has a clear transition between risk factor, however there is a lot of overlap. Different distance metrics in the weight calculation of the graph Laplacian may cause these algorithms to give more desirable results.

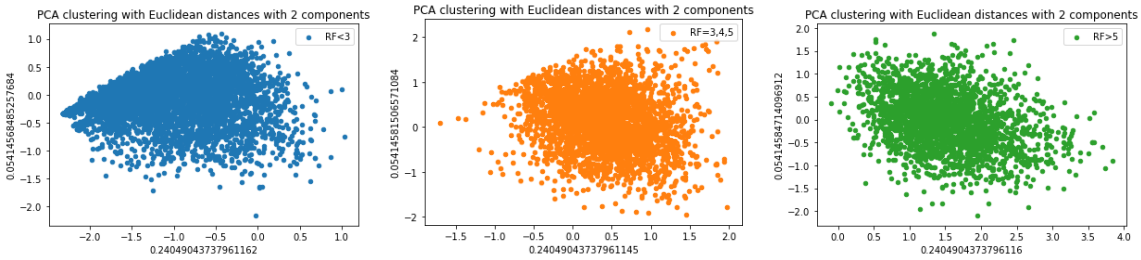

Figure 3: PCA on questions: overlap

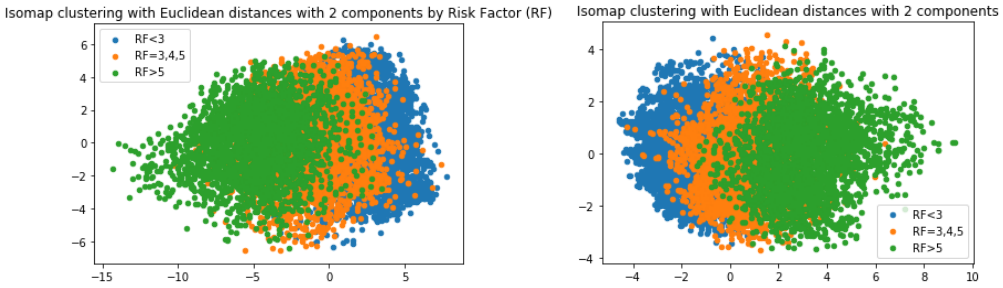

Figure 4: Isomap on questions. On the left, $k = 5$ and on the right, $k = 50$ for k-nearest neighbors.

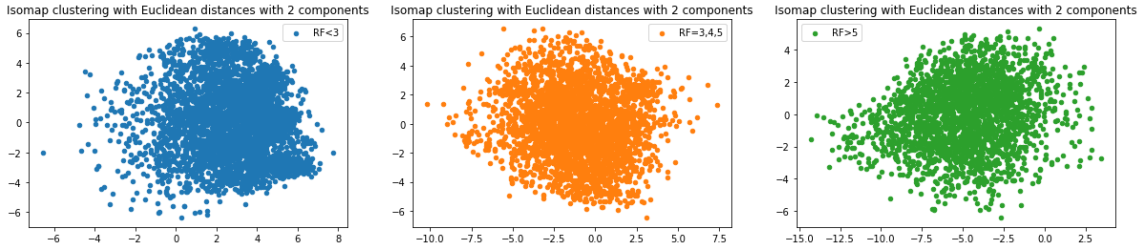

Figure 5: Isomap on questions: overlap

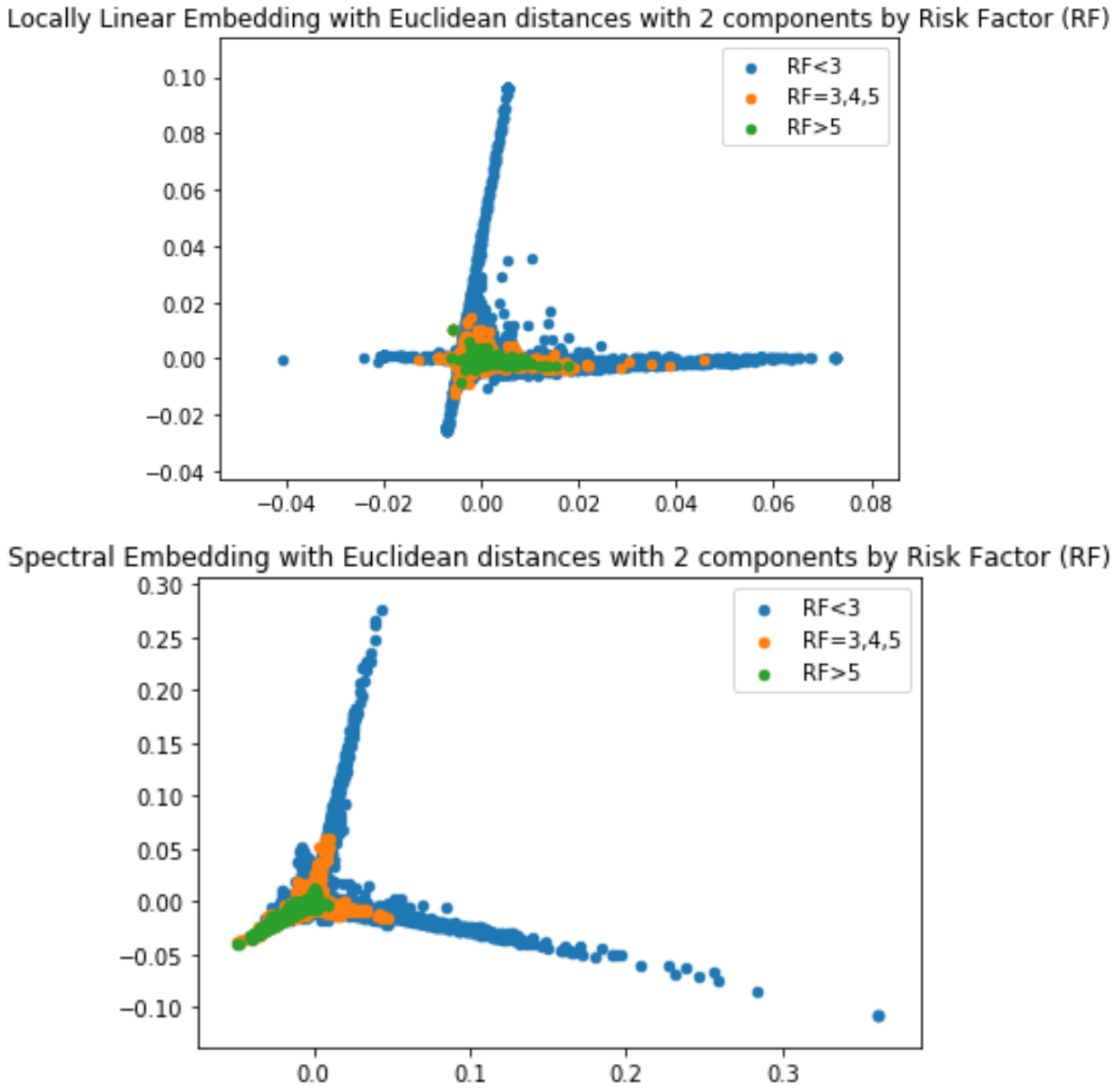

Figure 6: Other methods: Locally Linear Embedding, Spectral Embedding, and t-distributed Stochastic Neighbor Embedding on the questions