COSC 528 Project 2: Dimensionality Reduction and Clustering

Elliot Greenlee

October 23, 2017

Abstract

This report covers the design and implementation of dimensionality reduction and two parametric clustering methods in order analyze infant mortality data. Principal component analysis by singular-value decomposition, k-means clustering, and the expectation-maximization algorithm for Gaussian clusters are used on infant mortality rate data by country from 1800 to 2015 in order to aid visualization, find trends, and cluster the countries. From this analysis, it can be observed that countries fall into two categories: those that have low infant mortality either through large improvement or continual success, and those that have higher infant mortality with little improvement.

1 Introduction

As the sizes of databases grow, computational methods to extract useful information from disorganization are needed. Methods for the analysis of input data can take two forms: either supervised or unsupervised. In supervised learning, possible or actual classes for the data are known. In unsupervised learning, the classes of samples are not known, and the number of appropriate classes must be determined. Clustering is used to solve this problem, finding groups of similar samples in the data as in figure 1 from researcher Kadir Peker.

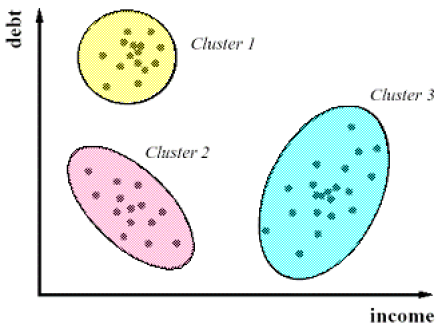


Figure 1: Income vs. debt cluster example

In this case, the discrimination is easy. With more data, more dimensions, and closer clusters, the problem becomes more challenging, and more appropriate for a computer to solve. Between the features of samples is a calculated measure of similarity, such as Euclidean distance. This metric can be used to compare samples, and to find clusters of similar samples [3].

The objective of this project is to implement Principal Component Analysis by Singular Value Decomposition, k-means clustering, and expectation maximization for Gaussian clusters and apply them to infant mortality data in order to provide trend and cluster analysis. Sections 2 and 3 provide more context for the algorithms used and section 4 reports the results of analysis and gives more context for the dataset used. Section 5 summarizes the work.

2 Dimensionality Reduction

Although in theory a greater number of statistically independent features should reduce further and further the error, in practice additional features often lead to worse performance. This occurs either

when the wrong model is used, or because the training sample is not infinite, so the distributions are not accurately estimated. This leads to problems with overfitting, where the model works well on the training data but poorly on the testing data, and complexity, where the model takes longer to run at $O(d^2n)$ where d is dimension and n is number of samples. The solution to these problems is dimensionality reduction. Linear methods are used because of their simplicity, projecting high dimensional data onto a lower dimensional space.

One common method is principal component analysis, with the objective of minimizing information loss in a least-squares sense. Principal component analysis attempts to reduce the data while minimizing information loss. One might imagine the major and minor axis of a set of two dimensional elliptical data as in figure 2. Principal component analysis would choose the major axis as the projection direction to preserve information. Because there is no consideration of class information, it is an unsupervised method [3].

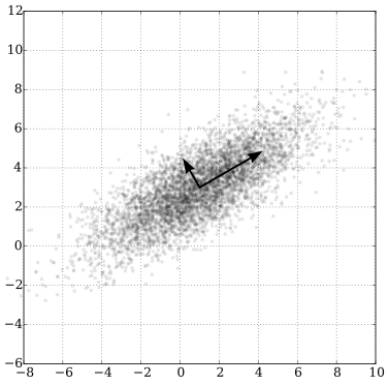


Figure 2: An example of the first two principal components on a graph.

One option for principal component analysis is singular value decomposition, which can decompose any Nxd matrix X into components VAW^T where V is the NxN matrix of the eigenvectors of XX^T in its columns, W is the dxd matrix of the eigenvectors of X^TX in its columns, and A is the Nxd matrix of the q singular values on its diagonal. For A, $q = \min(N, d)$, where the values are square roots of the nonzero eigenvalues of both XX^T and X^TX. The principal component P matrix of shape Nxd can be constructed as $P = (W^TX^T)^T$, where each row of P corresponds to one of the N instances. To reduce, take the first k columns of P [1].

Once principal component analysis has been performed, the number of components k to use remains a problem. The goal for choosing a point at which to cut of extraction is to encompass all non-trivial variance; however, what constitutes non-trivial depends on the circumstances. To this end, we use two methods for choosing k: the scree plot and percent variance. The scree plot plots the eigenvalues in descending order, and k should be chosen at the "elbow" of the graph, where the values begin to decline more slowly. Percent variance v can be calculated as $v_i = \frac{s_i^2}{\sum_n s_n^2}$ where s is a vector of the n singular values [2]. Generally, a total percent variance cutoff for the problem is chosen, say 99% is chosen, and when the first k component variances sum to that variance cutoff, stop. Obviously, both these tests are highly dependent on problem circumstance.

3 Clustering

Clustering methods can be separated into parametric and non-parametric groups. In parametric methods, a model for the data, such as Gaussian clusters, is assumed. The number of classes to attempt is input at the start of computation. This requires a brute force implementation on each number of classes, and can be time consuming on large sample sizes. In non-parametric methods, no prior knowledge is used. Instead, another measure of associativity is used for computation. K-means and EM for Gaussian clusters are both parametric methods. In fact, k-means can be thought of as a simplified version of EM using a single covariance value.

3.1 K-Means

K-means is a centroid-based clustering technique, where clusters centers are represented by a structure identical to the individual samples. The algorithm looks for the optimal cluster placement such that the squared distances to each sample in the cluster is minimized. The algorithm begins by creating arbitrary cluster centroids, and then assigning samples to the nearest centroid.

$$b_i^t = \left\{ \begin{array}{ll} 1 & \text{if } ||x^t - m_i|| = \min_j ||x^t - m_j|| \\ 0 & \text{otherwise} \end{array} \right\}$$

The centroid is then recalculated as the mean of the assigned samples.

$$m_i = \frac{\sum_t b_i^t x^t}{\sum_t b_i^t}$$

Samples are reassigned, and if any sample classification has changed, then the centroid is recalculated and the process begins again [1].

3.2 Expectation Maximization for Gaussian Clusters

EM attempts to look for component density parameters that iteratively maximize the likelihood of the sample. This work uses Gaussian density as the parametric model. Rather than calculating absolute membership as in k-means, instead likelihood is calculated for each sample t for cluster i in the expectation step as

$$h_i^t = \frac{\pi_i |S_i|^{-1/2} e^{-(1/2)(x^t - m_i)^T S_i^{-1} (x^t - m_i)}}{\sum_j \pi_j |S_j|^{-1/2} e^{-(1/2)(x^t - m_j)^T S_j^{-1} (x^t - m_j)}}$$

In the maximization step, the mean and covariances for cluster i are calculated as

$$m_i = \frac{\sum_t h_i^t x^t}{\sum_t h_i^t}$$

$$S_i = \frac{\sum_t h_i^t (x^t - m_i)(x^t - m_i)^T}{\sum_t h_i^t}$$

The original means and covariance matrices are calculated using the clusters initialized by k-means [1].

4 Results

The data for this report is a set of under five mortality rates per 1000 deaths by country from 1800 to 2015. Values given are real valued or missing. There are 275 total countries represented, of which 66 are missing all data, and 25 are missing partial data, usually from earlier years. All 91 of these entries have been removed from analysis, given a lack of prior knowledge in this subject area that could aid prediction.

Mean, minimum, and maximum deaths for 1800 and 2015 are shown in table 1. The greatest improvement from 2015 to 1800 is 500 in South Korea, and the least improvement is 293 in Chad. The top five greatest and least improving countries are shown in table 2. Figure 3 shows a plot of the average under-five deaths from 1800 to 2015, from which can be observed that, encouragingly, the number of under-five deaths has decreased over the last 216 years. This information was provided as a .csv and Excel file without source as part of COSC 528, Fall 2017 at the University of Tennessee, Knoxville [5].

Year	Mean	Minimum (Country)	Maximum (Country)
1800	425	322 (Belgium)	540 (Yemen)
2015	33.0	157 (Angola)	1.9 (Luxembourg)

Table 1: Mean, minimum, and maximum under-five deaths per 1000 live births for 1800 and 2015.

Country	Reduction in Infant Mortality
South Korea	500
Yemen	493
Kuwait	492
Iran	491
Nicaragua	487
Zimbabwe	316
Somalia	315
Central African Republic	313
Lesotho	310
Chad	293

Table 2: Countries showing the most and least improvement from 1800 to 2015 in infant mortality.

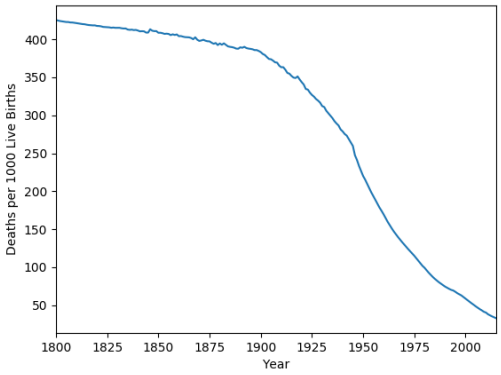


Figure 3: Average under-five deaths per 1000 live births across all countries from 1800 to 2015.

Different variables can all have wildly different ranges and variances, meaning that weights for different variables are not consistent in meaning. In many machine learning methods, such as clustering, this can have a detrimental effect on the ability to learn a model. For this project, z-normalization is used to transform the inputs into outputs with a mean close to zero and a standard deviation close to one. This is done for each input x_i using $z_i = \frac{x_i - \mu_i}{\sigma_i}$, where μ_i is the mean and σ_i is the standard deviation of the data [1].

4.1 Dimensionality Reduction

An appropriate value of k for dimensionality reduction to the k principal components using singular value decomposition is chosen by observation of both the scree plot and a plot of k versus the sum of percent variance up to k, both shown in figure 4. $k = 5$ is observed as the elbow of both graphs, and is chosen for reduction.

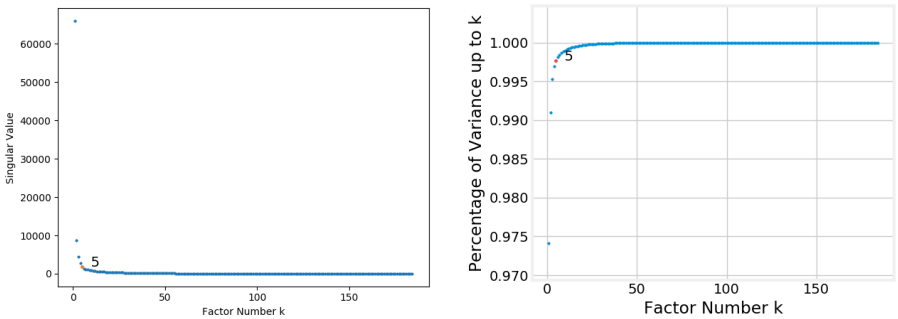


Figure 4: Scree plot of the singular values in descending order and k versus the sum of percent variance up to k.

Additionally, $k = 2$ is chosen so the results can be visualized graphically, as in figure 5. The five countries with the largest values and smallest values for each of the first two principal components are shown in table 3. Given prior knowledge of these countries, principal component one is observed as correlating negatively with typically underprivileged countries, and positively with infant mortality. Principal component two is more difficult to see a pattern, but is observed as correlating negatively with the amount of improvement made to infant mortality rates.

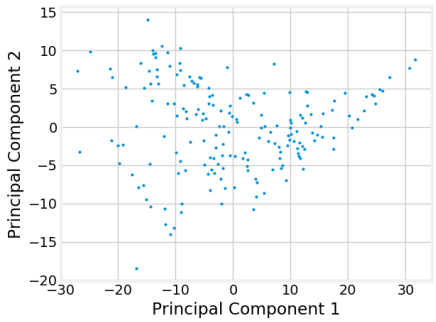


Figure 5: Countries plotted using the first two principal components.

Top Five	1	2	3	4	5
Highest PC 1	Norway	Sweden	Malta	Denmark	United Kingdom
Lowest PC 1	Sierra Leone	Yemen	Mali	Angola	Senegal
Highest PC 2	Niger	Chad	Guinea-Bissau	Guinea	Mali
Lowest PC 2	Barbados	South Korea	Fiji	Kuwait	Kazakhstan

Table 3: Countries with the top five highest and lowest principal components one and two.

4.2 Clustering Success

For this set of data there is no known classification of different countries. Because of this, accuracy can not be used as a metric for success. Instead, a measure of cluster separation called the Dunn Index is used. According to the index, clusters are said to be compact separated if and only if any pair of points within a cluster are closer together than any pair of points between two clusters, calculated as the minimal inter-cluster distance divided by the maximal intra-cluster distance [4].

$$DI_m = \frac{\min_{1 \le i < j \le m} \delta(C_i, C_j)}{\max_{1 \le k \le m} \Delta_k}$$

There are multiple possibilities for calculating inter- and intra-cluster distance, but for this project the methods with the lowest runtime complexity are chosen, both $O(n)$.

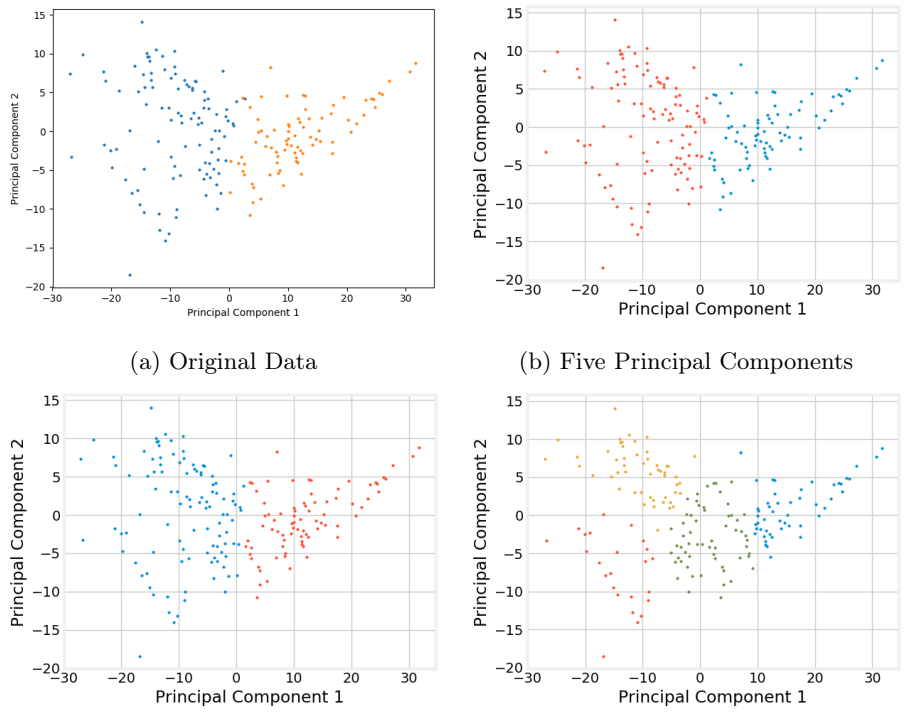
$$\Delta_i = \frac{\sum_{x \in C_i} d(x, u)}{|C_i|}, \quad \delta(C_i, C_j) = d(u_i, u_j), \quad u = \frac{\sum_{x \in C_i} x}{|C_i|}$$

4.3 K-means Clustering

The k-means clustering algorithm was run on the original and principal component reduced data at various values of k to find k clusters, and the Dunn Index scores were compared for each. The results of this testing can be found in table 4, with the highest values highlighted in red. For most data, two clusters showed the greatest match to the data, with four clusters also matching quite well. Graphs of the best matching clusters are shown in figure 6

Data	2	3	4	5	6	7	8	9
Original Data	2.4	1.8	1.5	1.0	1.3	1.1	0.68	1.1
5 Principal Components	2.2	1.6	1.7	1.7	1.3	1.3	1.2	1.0
2 Principal Components	2.5	1.7	2.3	2.0	1.6	2.1	1.3	1.9

Table 4: Dunn index scores for the original data and the data after being reduced to five and two principal components for each number of clusters using k-means. The highest values are shown in red, and are graphed.



(a) Original Data

(b) Five Principal Components

(c) Two Principal Components

(d) Two Principal Components

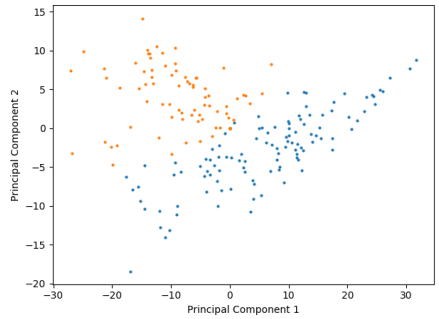
Figure 6: Results of k-means clustering on various data. Graphs (a)-(c) show two clusters. Graph (d) shows four clusters.

4.4 Expectation Maximization for Gaussian Clusters

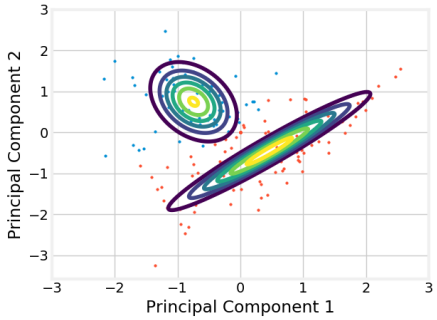
Similar testing was done using EM for Gaussian clusters. These results suffered from issues calculating the determinant of the ill-conditioned original data matrix. Instead, results from the five and two principal component cases are shown in table 5 and figure 7. Additionally, the Gaussian distributions during the final step of EM are plotted on the two principal component data using two clusters in figure 7 (b). From these results, it appears that there are two categories of country in the data.

Data	2	3	4	5	6	7	8	9
5 Principal Components	0.12	0.065	0.093	0.074	0.070	0.070	0.028	0.053
2 Principal Components	0.14	0.12	0.060	0.072	0.035	0.033	0.042	0.029

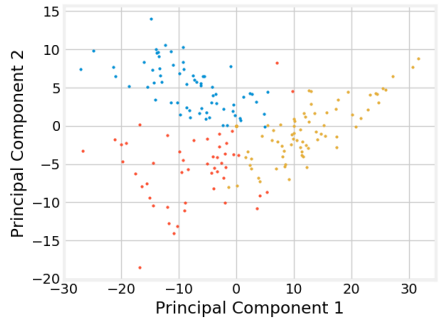
Table 5: Dunn index scores for the original data and the data after being reduced to five and two principal components for each number of clusters using expectation maximization. The highest values are shown in red, and are graphed.



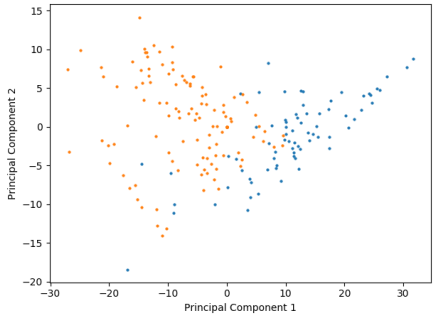
(a) Two Principal Components



(b) Two Principal Components



(c) Two Principal Components



(d) Five Principal Components

Figure 7: Results of EM clustering on various data. Graphs (a) and (b) show two clusters, with the Gaussian distribution in the final step of EM plotted on (b). Graph (c) shows three clusters. Graph (d) shows two clusters when using five principal components.

5 Summary

Dimensionality reduction by singular value decomposition to the principle components, k-means clustering, and expectation maximization for Gaussian clusters were implemented. Each method was applied to visualization and clustering of under-five infant mortality rates per 1000 live births. Investigation was done to determine the best number of clusters. It was determined that two clusters best captures the structure of the data.

The objectives of this project were to investigate principal component analysis and parametric methods for clustering on an unsupervised analysis problem. From this I learned how to implement the methods, and how to visualize and analyze unfamiliar data with many features. I also learned more about clustering metrics. In the future, more background investigation into the dataset, as well analyzing using other clustering methods and metrics would be interesting.

References

[1] Ethem Alpaydin. *Introduction to Machine Learning Third Edition.* The MIT Press, 2014.

[2] Raymond B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966. PMID: 26828106.

[3] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification Second Edition.* Wiley, 2001.

[4] J.C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separate clusters. *Cybernetics and Systems*, 3(3):32–57.

[5] Bruce MacLennan. Under 5 mortality rates 1800 to 2015, 2015. http://web.eecs.utk.edu/~mclennan/Classes/425-528/.

6 Appendix

The code for this project can be found on GitHub at https://github.com/LambentLight/528p2-cluster-pca. Please contact me with any questions or considerations.