Evaluating PCA For Dimensionality Reduction

Matthew Sun

1 Abstract

The purpose of this assignment is to use perform three different dimensionality reduction methods on a labelled data set and compare them using the Davies-Bouldin (DB) index and visual observation of the plotted reduction. The first method is to find the pair of data columns that produce the lowest DB index. The second was using PCA on the zero-mean data matrix and scoring the reduction. The third was using PCA on the standardized data matrix and scoring the reduction. It was found that the third method produced the best dimensionality reduction based on it having the lowest DB index of 0.6392. The first method ranked second with a DB index of 0.7875. The second method ranked last with a significantly higher DB index of 1.5148. Visual observation of the plotted reduction found that the second method had more data points within close proximity of other points with different labels and outliers in some clusters which contributed to a higher DB index, indicating a poor dimensionality reduction.

2 Introduction

The objective of the assignment was to perform, on a labelled data set containing samples of wine, three different types of dimensionality reduction methods: pair of variables with the lowest Davies-Bouldin (DB) index, Principal Component Analysis (PCA) on a zero-mean data matrix, and PCA on a standardized data matrix, then evaluate and compare their effectiveness. Each of the methods will provide a two-dimensional approximation of the matrix to which they were applied. Zero-mean and standardized matrices can be computed by finding $X = [A - [\vec{1}\bar{A}]]D_A^{-1}$, where $A \in \mathbb{R}^{m\times n}$ is the data matrix, $\bar{A}$ is the mean row matrix containing entries $\bar{a} := \frac{1}{m}\sum_{i=1}^{m} a_i$, D_A is the diagonal standard deviation matrix of A with entries $d_{ii} = \sigma_{ai} := \frac{||\vec{a}_i - \vec{1}\bar{a}_i||}{m-1}$. Finding a standardized data matrix is equivalent to treating every column as a normal distribution and assigning every entry its corresponding Z-score.

PCA is a dimensionality reduction technique used to approximate a given data set. One method of performing PCA is to use the singular vectors of the right transpose product V in the singular value decomposition (SVD) of a zero-mean data matrix $M = U\Sigma V^T$, $M \in \mathbb{R}^{m\times n}$ as the principle component(s). SVD is a factorization of a rectangular matrix that can be considered the generalization of eigendecomposition for square matrices and describes the "eigenvectors" of a rectangular matrix. Any square matrix $A \in \mathbb{R}^{m\times m}$ can be decomposed into $A = Q\Lambda Q^{-1}$ where Q is a square matrix whose columns are the eigenvectors of A and Λ is a diagonal matrix with entries $\Lambda_{ii} = \lambda_i$, where λ_i is the corresponding eigenvalue. Similarly, SVD factorizes any rectangular matrix $X \in \mathbb{R}^{m\times n}$ into $X = U\Sigma V^T$. U is an orthogonal matrix whose columns are the eigenvectors of $[XX^T]$ which form an orthonormal basis for the data space $\mathbb{R}^m$, also known as the left singular vectors. If X is a full rank matrix that has fewer rows than columns, meaning $m < n$. In that case, $[XX^T]$ is a square symmetric matrix of size $m \times m$. If X is a "tall thin" matrix, meaning $m > n$, or X is square and rank deficient, then XX^T is also a diagonalizable square positive semi-definite matrix with size $m \times m$. In either case, the decomposition can be written as $[XX^T] = U\Lambda U^T$, where Λ is a rectangular diagonal matrix with entries $\Lambda_{ii} = \lambda_i$ for i in range 1 to r, where r is the rank of the matrix. The entries are arranged in descending order such that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r$. These entries are called the singular values of X. V is an orthogonal matrix whose columns are the eigenvectors of $[X^TX]$ which form an orthonormal basis for the weight space $\mathbb{R}^n$, also known as the right singular vectors. If X has more rows than columns then $[X^TX]$ is a square symmetric positive definite matrix of size $n \times n$ that can be decomposed to $[X^TX] = V\Lambda V^T$ where Λ is constructed the in the same fashion as described previously. The first r columns of V, where r is the rank of X, form a basis for the weight space of X. The non-zero entries λ_j of $[AA^T]$ and $[A^TA]$ are identical. This can be shown by factorizing each of the products with SVD and using the fact that the transpose of an orthogonal matrix is its inverse:

$$[AA^T] = [U\Sigma V^T][U\Sigma V^T]^T \tag{1}$$
$$= U\Sigma\Sigma^T U^T \tag{2}$$
$$= U\Lambda U^T \tag{3}$$
$$[A^T A] = [U\Sigma V^T]^T[U\Sigma V^T] \tag{4}$$
$$= V\Sigma^T\Sigma V^T \tag{5}$$
$$= V\Lambda V^T \tag{6}$$

This also shows that Σ is a rectangular diagonal matrix with entries $\Sigma_{ii} = \sigma_i := \sqrt{\lambda_i}$, where each λ_i an eigenvalue of $[AA^T]$ and $[A^T A]$. Additionally, Σ must satisfy these properties: each entry $\sigma_j \in \mathbb{R}$, $\sigma_j > 0$ for $j \leq r$, $\sigma_1 \geq \sigma_2 \geq ... \geq 0$, and r is the rank of X. Σ may have diagonal entries of 0 for $i > r$.

PCA can also be performed by calculating the sample covariance $B = \frac{M^T M}{m-1}$ or scatter matrix $S = M^T M$ which can be thought of as the weighted covariance, where M is a zero-mean data matrix and m is the number of observations. Each eigenvalue of $M^T M$ corresponds to the proportion of the variance that is associated with each eigenvector. A larger eigenvalue implies that the corresponding eigenvector captures more of the variance in the data. Suppose that M has n variables. Dimensionality reduction reduces a linear problem to $p < n$ dimensions. One method for doing so is calculating the first p scores $Z_p = MV_p$, where $p \leq$ is the rank of M. This provides an optimal approximation for the p-dimensional vector space of the zero-mean data matrix. Factorizing $[M^T M] = V\Sigma^2 V^T$ shows that the principle component(s) of PCA are the right singular vectors from V of the SVD [BDG09].

The Davies-Bouldin (DB) index is a method for measuring the effectiveness of a clustering by describing how far each cluster is away from each other, and how scattered are the elements of a set. Suppose there are two clusters, S_1, S_2 with centroids $\vec{g}_1$, $\vec{g}_2$ respectively, every data vector $\vec{x}_i$ is in either S_1 or S_2. The distance between the two clusters can be measured using the norm of the difference of the centroids $||\vec{g}_1 - \vec{g}_2||$. The dispersion within a set is measured using the mean distance of every point to its centroid: $d_1 = \frac{1}{m_1}\sum_{i=1}^{m_1} ||\vec{x}_i - \vec{g}_1||$, $d_2 = \frac{1}{m_2}\sum_{j=1}^{m_2} ||\vec{x}_j - \vec{g}_2||$, where m_1, m_2 are the number of observations in each set. The DB index is defined as $I := \frac{d_1+d_2}{||\vec{g}_1-\vec{g}_2||}$, this index minimizes the dispersion while maximizing the distance between centroids. A lower DB index is an indicator of a more effective clustering. k-means clustering is an unsupervised learning algorithm that finds natural clusters in a given non-empty data set $X \in \mathbb{R}^m$ that has m members $\vec{x}_j$. A cluster S_i can be defined as a set with $m_i > 0$ members that is part of a partition of X such that $S_i \subset X$. Alternatively, a cluster S_i is the set with a centroid $\vec{g}_i \in X$ such that for any distinct centroid $S_i = \{\vec{u} \in X : ||\vec{u} - \vec{g}_i|| - ||\vec{u} - \vec{g}_k||\}$, where $\vec{g}_i \neq \vec{g}_k$. A common implementation of the k-means clustering algorithm is to randomly initialize k centroids, then while the clustering has not converged, each $\vec{x}_j \in X$ is assigned to the index i of the nearest centroid $\vec{g}_i$, then for each index i, calculate a new $\vec{g}_i$ for all $\vec{x}_j$ that has index i. Note that there are many ways to measure convergence and different measures may affect the resulting clustering.

The scientific question to be explored is: which of the three dimensionality reduction methods described previously produces the best clustering and thus the best dimensionality reduction when using the entries of the cultivar vector as the cluster indices? The effectiveness of each method will be evaluated using the DB index of the dimensionality reduction and visual observation of the plotted reductions. A lower DB index indicates a more effective clustering and thus a more effective dimensionality reduction.

3 Methods

Let $A \in \mathbb{R}^{m\times n}$ be the data matrix of the given data set with a label vector $\vec{l}$ that contains cultivar numbers 1, 2, 3. The pair of variables (column vectors) in A that resulted in the lowest DB index was found by iterating through every variable $\vec{a}_i$ then computing the DB index using $\vec{l}$ with each $\vec{a}_j$ where j is in the range $i + 1$ to n. Each computed DB index is stored along with the indices of the pair of variables that

produced the index. When all variables have been paired with each other, the lowest DB index and the corresponding indices of the columns are stored. The columns at those indices of that are the dimensionality reduction for the first method.

Let $M \in \mathbb{R}^{m \times n}$ be the zero-mean data matrix. PCA as dimensionality reduction was performed by finding the right singular vectors of $M = U\Sigma V^T$. The SVD was found by using the Matlab function "svd". Taking the first two columns of V and computing $M[\vec{v}_1\vec{v}_2]$ provides an optimal two-dimensional reduction of M as the first two vectors will correspond with the largest singular values and thus account for the most variance in M. Let the score $Z_M = M[\vec{v}_1\vec{v}_2]$, the DB index for the second method was calculated using Z_M and $\vec{l}$. The calculations for the third method are identical to the second except SVD is performed on the standardized matrix X which will result in the two-dimensional reduction $Z_X = X[\vec{v}_1\vec{v}_2]$. The reduction of each method was then plotted with labels $\vec{l}$.

The DB indices produced by each method were used to evaluate the effectiveness of the dimensionality reductions and compare methods with one another. The scatter plots of the reductions were assessed by visual observation to complement the measurement provided by the DB indices.

4 Results

Table 1: Left column contains each method, the middle column contains the respective Davies-Bouldin (DB) index to four digits of precision, right column contains the indices of the pair of the variables that produced the best dimensionality reduction based on the DB index.

Method	DB Index	Variables
Data Columns	0.7875	[1, 7]
Raw PCA scores	1.5148	N/A
Standardized PCA scores	0.6392	N/A

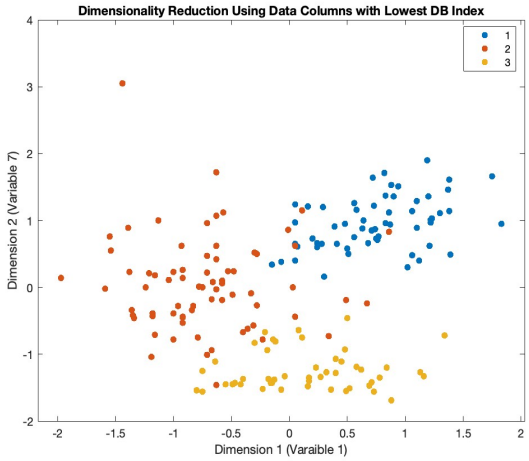

Figure 1: Scatter plot of dimensionality reduction using the pair of data columns that produced the lowest DB index with the cultivar vector as the labels. Cultivar 1 in blue, 2 in red, 3 in yellow.

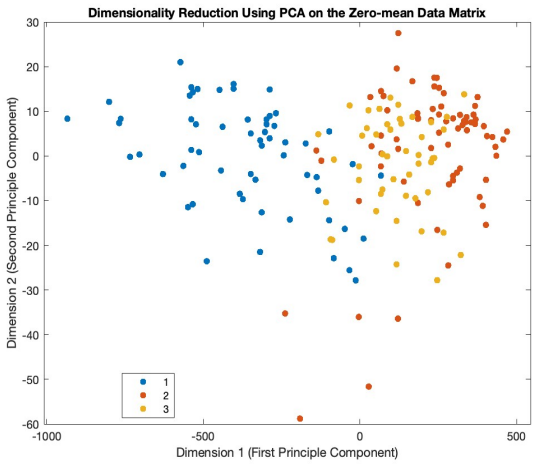

Figure 2: Scatter plot of dimensionality reduction using PCA on the zero-mean data matrix with the cultivar vector as the labels. Cultivar 1 in blue, 2 in red, and 3 in yellow.

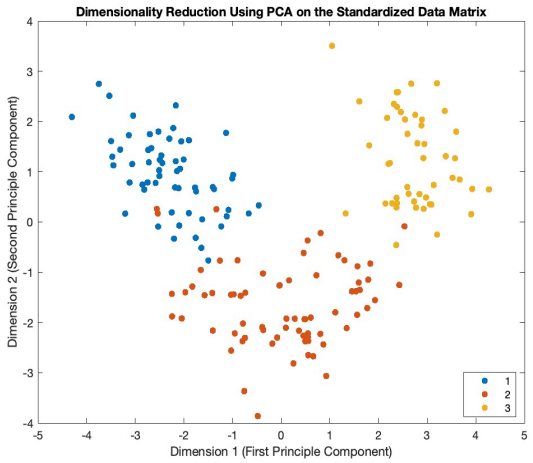

Figure 3: Scatter plot of dimensionality reduction using PCA on the standardized data matrix with the cultivar vector as the labels. Cultivar 1 in blue, 2 in red, and 3 in yellow.

5 Discussion

For the given data set containing samples of wine of three different cultivars, performing dimensionality reduction using PCA on the standardized data matrix produced the best clustering based on the finding that its DB index using the cultivar numbers as the labels was the lowest (best) amongst the three methods, at 0.6392. It was found that using variables 1 and 7 produced the lowest DB index when using pairs of data columns to perform dimensionality reduction; this method yielded the second lowest DB index at 0.7875. Performing dimensionality reduction using PCA on the zero-mean data matrix produced the highest (worst) DB index at 1.5148. Visual observation of the plotted approximations would agree with the evaluation provided by the DB indices of the methods. The DB index of PCA on the zero-mean data matrix (1.5148) differs the most from the other two methods meaning it is significantly worse. Observe the top right quadrant of Figure 2, there are many data points from clusters 2 and 3 that are in close proximity to each other. Such a phenomenon would decrease the distance between the centroids of clusters 2 and 3. It can also be observed from the bottom half of Figure 2 that there are data points from cluster 2 that lie further from the rest which will increase the mean distance between data points of cluster 2 to its centroid. Both of these observations are indicators that the DB index for this method would be higher. In contrast, Figure 1 and 3 shows fewer of these indicators. There are fewer points within close proximity of points of another cluster and points of

the same cluster tend to be closer to each other. Although it can be seen from comparing Figures 1 and 3, Figure 1 has more points, in the center, that "overlap" visually with points from other clusters and cluster 2 has an outlier in the top left corner. Thus, it can be said that based on both the DB index and visual observation, applying PCA on the standardized matrix of the given data set produced the most optimal reduction to two dimensions amongst the three methods performed.

The labels used in the three methods were the original cultivar numbers from the given data set. By relabeling the data points of the reduction produced by each of the original three methods using the cluster indices assigned by the k-means clustering algorithm, natural clusters in each of the reductions were revealed. It was observed that the labelling and DB indices were different than the indices calculated using the original cultivar numbers. The DB index using the two best variables went from 0.7875 to 0.7090, applying PCA on the zero-mean data matrix went from 1.5148 to 0.5330, and applying PCA on the standardized data matrix went from 0.6392 to 0.5979. Note that the new DB indices will not be consistent run-to-run and the differences between the original DB index and the DB index using the labels produced by the k-means algorithm are not necessarily an indicator of the effectiveness of the dimensionality reduction methods. Rather, it is an indicator of the clustering quality of the reduction. In other words, if the data were to be represented in two dimensions using the approximation provided by each of the methods, the indices from k-means clustering are a way to naturally relabel the data points, and the new DB indices provide a measure of the quality of the relabeling. This is especially applicable to using data columns to perform dimensionality reduction as it could be stated that the new labels from the k-means clustering algorithm are directly related to which variables were used. Furthermore, reducing to only two dimensions is a limitation to measuring the effectiveness of the methods. Further testing should be conducted where the data is reduced to three or four dimensions and repeating the process of finding the DB indices with the original cultivar labels, then calculating the DB indices with the labels provided by the k-means clustering algorithm.

Using k-means to compare the effectiveness of dimensionality reduction methods can be useful as demonstrated in a 2019 study [FT19] where the researchers observed 87 dimensionality reduction methods including PCA, on data sets of mRNA and DNA methylation. They evaluated the methods based on four categories using five criteria. Dimension reduction quality was evaluated using the Local Continuity Meta-Criterion (LCMC) and measured as local, global, and average values. Speed was evaluated using an approximation of the runtime; the authors acknowledge that runtime depends on a variety of factors but the estimation provides an indication of speed. Sensitivity was measured by varying one parameter each run while keeping all other parameters constant, recording the local dimension reduction quality scores, then finding the standard deviation for the collected values. Clustering quality was evaluated by performing k-means and computing the Normalized Mutual Information (NMI) criterion. It was found that PCA was in the top five for global dimensionality reduction. The study also stated that when considering the qualities of a good global dimensionality reduction which are high reduction quality, lower sensitivity, and speed, PCA ranked well in all factors and "that despite the recent development of many DR techniques, considering the three factors accuracy, speed and robustness, PCA is still considered state-of-the-art". However, PCA was not amongst the top performers in the clustering quality category as it fell in the 40_{th}-60_{th} percentile.

For the given data set it was found that performing PCA on the standardized matrix provided the most optimal dimensionality reduction to two dimensions based on its DB index using the cultivar numbers as labels and visual observation. Using the pair of variables that provided the lowest DB index ranked second and using PCA on the zero-mean data matrix ranked last. Visual observation of the plotted reductions agrees with the differences amongst the DB indices. Relabeling the reductions using the cluster indices from the k-means clustering algorithm revealed further information about the clustering quality of the reduction.

References

[BDG09] Daniel P. Berrar, Werner Dubitzky, and Martin Granzow. *A practical approach to microarray data analysis*. Springer, 2009.

[FT19] Hadi Fanaee-T and Magne Thoresen. "Performance evaluation of methods for integrative dimension reduction". In: *Information Sciences* 493 (2019), pp. 105–119. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2019.04.041. URL: https://www.sciencedirect.com/science/article/pii/S002002551930355X.