Linear Data Analysis
PCA - Matrix Algebra and Dimensionality Reduction

Cain Susko

Queen's University
School of Computing

March 8, 2022

a Revisiting the PCA

this section covers the PCA and will reiterate its axes and scores.

Data Matrix Recall that, within a data matrix, a variable (column) is a real number with a type. Type has meaning but is not categroical. Thus, an observation is a ordered list of valuations.
With this clarified we can then take the differences (e.g. from the mean) and the score as a weighted sum. we shall then gather this data into the matrix $A \in \mathbb{R}^{m \times n}$

Vector Spaces: Dimensions the vector spaced of a full-rank $A \in \mathbb{R}^{m \times n}$ has the vector spaces:

- Column Space, the general form of the columns within A
- Row Space, the general form of the rows within A

Note: the class data generally has a much greater m than n: $m >> n$
The dimensionality reduction is done by only using the parts of the right Singular Martix that are within $\mathbb{R}^n$:

$$\mathbb{V} \subset \mathbb{R}^n$$

An example of this would be to reduce the grades of 3 quizzes into 1 score.

Terminology the PCA of A has the following related variables

$$M = A - \vec{1}\bar{A} \qquad \text{Zero Mean Matrix}$$

$$B = \frac{M^\top M}{m-1} \qquad \text{Covariance Matrix}$$

$$\vec{v}_1, \vec{v}_2, ... = eigenVectors(B) \qquad \text{Loading Vectors}$$

$$\lambda_1, \lambda_2, ... = eigenValues(B) \qquad \text{Latent Variables}$$

b The Scatter Matrix of Variables for PCA

We will expolre the scatter matrix of the data
$A \in \mathbb{R}^{m \times n}$ and $M = zeroMean(A)$. The scatter matrix of A would be:

$$S =^{def} M^\top M$$

where S is positive semidifinite.
S has same eigenvectors as the convariance matrix B $(\vec{v}_1, ...)$ and has the same scaled eigenvalues as B such that: $(m-1)\lambda_i$.
Thus, S is the **Weighted Covariance**.

Recall The Spectral Throrem where the eigenvectors $\vec{v}$ of a symmetric matrix are a ortho-normal basis for all of the vectors with m entries $\mathbb{R}^m$. Thus, the loading vectors $\vec{v}_i$ are the coordinate axis for $\mathbb{R}^m$

Algebra we get 'better' coordinates from using algebra rather than the PCA. The Geometry that is derived from this algebra can be used to describe the coordinate axes.

c PCA as Matrix Approximation

given the zero mean matrix $M \in \mathbb{R}^{m \times n}$ which has the form $M = U\Sigma V^\top$

Rank-p Approximation an approximation of the matrix M with rank p is equal to:

$$M = \vec{u}_1\sigma_1\vec{v}_1^\top + \vec{u}_2\sigma_2\vec{v}_2^\top + ... + \vec{u}_p\sigma_p\vec{v}_p^\top$$

Recall that:

$$\vec{v}_i^\top \vec{v}_i = 1$$
$$\vec{v}_i^\top \vec{v}_{j \neq i} = 0$$

and thus the PCA Score is:

$$\vec{z}_1 = M\vec{v}_1 \equiv \sigma_1\vec{u}_1$$

Note: This is the score only for the vector $\vec{v}_1$.

Equivalencies from this information, we can say that that following values are equivalent

$$\text{First PCA Score } \vec{z}_1 \equiv \text{ approximation } \sigma_1 \vec{u}_1$$

Additionally, the following approximations are equivalent

- Approximate M as a rank-p Matrix using $Z_p = M_p V_p$
- Approximate the column space of M as either Z_p or U_p

thus, if $p = 2$ then:

$$Z = [\sigma_1 \vec{u}_1 \quad \sigma_2 \vec{u}_2]$$

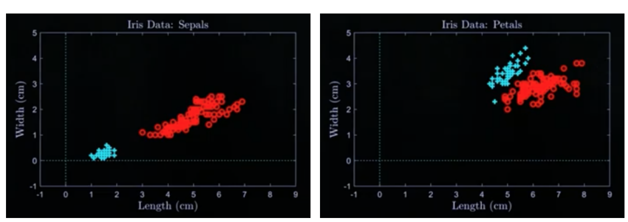
and equivalently:

$$Z = [\vec{u}_1 \quad \vec{u}_2]$$

where both score matricies span the space U_2.
thus this shows us that approximating the data using the scores Z and approximating the data using the left singular vector U result in the same conclusion as they span the same vector space.

d PCA as Dimensionality Reduction

we shall explore Dimensionality Reduction in terms of the PCA. Using Fisher's Iris Data (petals and sepals) and 2 labels (beach-head Iris, everything else)



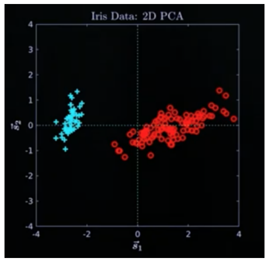
we can see that the Sepal Data is aligned along a skewed axis.
Using PCA we can reduce the dimensionality of this data. first, we will define our zero mean matrix as:

$$M = [\vec{m}_1 \vec{m}_2 \vec{m}_3 \vec{m}_4]$$

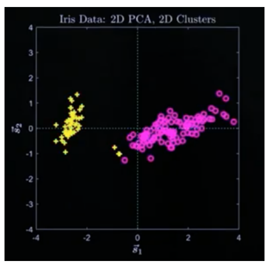
where columns 1-4 are the zero mean vectors of: 1; petal length, 2; petal width, 3; sepal length, 4; sepal width.
We will then reduce M using the PCA such that we score each variable within M by doing $M\vec{v}_i$ where the vector v is the i^{th} loading vector; thus giving us:

$$\vec{S}_1, \vec{S}_2$$

We can then plot these 2 scores like so:



and if we then perform the kmeans algorithm with a squared euclidian norm on M we get the corresponding result:



and we find that kmeans and squared norm does not provide as good of a clustering PCA (observe the outliers in the kmeans plot)

Learning Summary

Students should now be able to:

- compute scores of data using the scatter matrix
- reduce the dimensionality of the data using the zeor-mean matrix and the loading vectors and apply clustering to the reduced data
- interperet some of the results visually