Linear Data Analysis
Principal Component Analysis

Cain Susko

Queen's University
School of Computing

March 2, 2022

a Introduction to the PCA

this note will explore the PCA, which is related to the covariance matrix, and singular vectors of difference. It can be used to reconstruct a matrix.
Principal analysis is the process of finding the the 'principal' ammount from which data varies from the 'average' data.
take, for example, a matrix of students and test scores $A \in R^{m \times n}$ where a variable is a column of scores for a given test; and an observation is a row of scores for all tests for a given person.

b PCA from Covariance Matrix

this section will expolre the covariance of data with respect to the PCA.
For a zero mean matrix M from the matrix A, the covariance of M is equal to:

$$cov(M) = \frac{1}{m-1} M^\top M = B$$

thus, B is the covariance matrix of A where each place $b \in B$ represents the variance between tests (from our student example above) i, j where i, j are the indicies of a. Note: this matrix is symmetrical and thus, $b_{ij} = b_{ji}$.
because B is both symmetric and positive semidifinite (values are ≥ 0) we know that it also has a Eigen-decomposition: $B = E\Lambda E^\top$.
This, thus, allows us to analyzie the covariance matrix B, from A, and make the following observations:

Principal Components eigenvectors of B
the principal components tell us how the variables in our data is related

Latent Variables eigenvectors of B and,
diagonal entries of the eigenvalue matrix Λ
the latent variables tell us how much each principal component contributed to the covariance.

Scores Products of zero-mean matrix M with the principal components

Note: the first eigenvector in E which is associated with the first eigenvalue in Λ is the best rank 1 approiximation to the covariance matrix B; which means that the aformentioned eigenvector is the best 1 dimensional approximation of the column space of M which is closely related to a 1 dimensional approximation the column space for A.

c PCA as Spectral Decomposition

givem the data $inR^{M\times n}$ with a covariance matrix $B \in R^{n\times n}$ where B is positive semidifinite with an eigen decomposition $B = E\Lambda E^\top$

$$\Lambda_1 = \begin{bmatrix} 75.60 & 0 & 0 \\ 0 & 4.82 & 0 \\ 0 & 0 & 0.17 \end{bmatrix}$$

$$E_1 = \begin{bmatrix} 0.54 & 0.79 & 0.29 \\ 0.48 & -0.58 & 0.66 \\ 0.69 & -0.22 & -0.69 \end{bmatrix}$$

where the first eigenvector is approximately $\vec{v}_1 = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.7 \end{bmatrix}$

Now, we can start to try and represent the matrix B using the Spectral Decomposition. recall, the covariance matrix is defined as

$$B = \frac{MM^\top}{m-1} = \frac{[U\Sigma V^\top]^\top [U\Sigma V^\top]}{m-1} = \frac{V\Sigma^\top U^\top U\Sigma V^\top}{m-1} = V\frac{\Sigma^\top\Sigma}{m-1}V^\top$$

thus, the loading vectors of B are the right singular vectors U of the zero mean data M. In practice, U may be quite large as it is $U \in R^{m\times m}$. to account for this in MatLab one could use what is called the 'economy' svd:

```
[U,S,V] = svd(M, 0)
```

the matrix `U` will have $U \in R^{m\times n}$ rows and columns, thus reducing the ammount of memory required from a computer.

PCA Scores of Data

First Loading Vector is the eigenvector which is associated with the largest eigenvalue in the covariance matrix (obtained through eigen decomposition). this 'weights' the i^{th} observation by $v_1, v_2, ..., v_n$ which results in:

$$v_1 m_{i1}, v_2 m_{i2}, ..., v_n m_{in}$$

the score fot component #1 is z_i. compute all scores for component #1 at once by doing:

$$\vec{z} = M\vec{v}$$

Note: the PCA error is Orthogonal error (it measures the distance from the model to the data orthogonal to the model, instead of vertically (along the y axis) from the model).

Learning Outcomes

- find the 'explained variance' from Latent Variables
- find the principal components of data.
- compute scores of data using Principal Compoinent Analysis