Principal Components Analysis: Using Math to Extract Underlying Structures of Iraqi Migration Survey Data

Jackie Lindstrom

Pacific Lutheran University

2023

Principal Components Analysis

PCA is a mathematical tool used to reduce the dimension of a data set in the hope of obtaining useful and understandable takeaways.

Linear Combination

Definition (Linear Combination)

A linear combination is any sum of vectors multiplied by constants, such as

$$\alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \alpha_3 \mathbf{u}_3 + \alpha_m \mathbf{u}_m = \sum_{j=1}^{m} \alpha_j \mathbf{u}_j \tag{1}$$

where $\mathbf{u}$ are vectors and α are constants. [2]

Linear Transformation

Definition (Linear Transformation)

A mapping f from $\mathbb{R}^n$ to $\mathbb{R}^m$ is **linear** if

$$f(a\mathbf{x} + b\mathbf{y}) = af(\mathbf{x}) + bf(\mathbf{y})$$

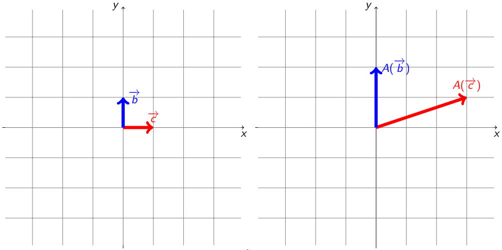
for all vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ and for all scalars a and b. [2]

Linear Transformation Example

Here, A is a square matrix representing some linear transformation.

$$A = \begin{bmatrix} 3 & 0 \\ 1 & 2 \end{bmatrix}$$

$$\mathrm{A}(\overrightarrow{b}) = \begin{bmatrix} 3 & 0 \\ 1 & 2 \end{bmatrix} * \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} (3*0)+(0*1) \\ (1*0)+(2*1) \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$



A can transform the vectors $\overrightarrow{b}$ and $\overrightarrow{c}$ to what we see on the right. Notice $\overrightarrow{b}$ stays on its span

Eigenvectors and Eigenvalues

Definition

Let **A** be any square matrix, real or complex. A number is an **eigenvalue** of **A** if the equation

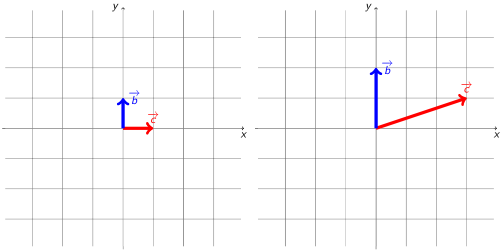
$$A\vec{v} = \lambda\vec{v} \tag{2}$$

is true for some nonzero vector $\vec{v}$. The vector $\vec{v}$ is an eigenvector associated with the eigenvalue λ. [2]

- left hand side of the equation: matrix vector multiplication
- right hand side: scalar vector multiplication

Visualizing Eigenvectors and Eigenvalues

$$A = \begin{bmatrix} 3 & 0 \\ 1 & 2 \end{bmatrix}$$



$$A\vec{v} = \lambda\vec{v} \tag{3}$$

Here, $\overrightarrow{b}$ is an eigenvector because it remains on its span (the y-axis) when transformed by A. However, $\overrightarrow{c}$ is not an eigenvector of A.

Factoring Matrices

Whenever we see a matrix written in the form

$$A = BC$$

we have a factorization of a matrix into the product of the two matrices B and C. [2]

Factorization can be useful to visualize and extract information from our matrix.

$$\begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}$$

Eigen Decomposition

One form of factorization is the eigen-decomposition.
Recall Equation 3

$$A\vec{v} = \lambda\vec{v} \tag{4}$$

Take U to be the matrix where each column is an eigenvector of A. Take Λ to be a diagonal matrix that stores the eigenvalues of A. Therefore, we can rewrite Equation 3 as:

$$AU = \Lambda U$$

Thinking in terms of a change of basis, we can write this as

$$A = U\Lambda U^{-1}$$

where U^{-1} is the inverse matrix of U.

Eigen Decomposition Continued

$$A = U \Lambda U^{-1}$$

$$A = \begin{bmatrix} 3 & 0 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} * \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} * \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

Therefore, the eigen-decomposition is a way of rewriting matrix A in a form that clearly shows what the eigenvectors and eigenvalues are.

- Eigenvalues: 3, 2
- Eigenvectors: [1,1], [0,1]

Singular Value Decomposition

- A generalization of the eigen-decomposition.
- Takes a matrix A and decomposes it into three factors
- Each factor gives useful information about the original matrix A
- Matrix A can be a non-square matrix (unlike the eigen-decomposition) [1].

Singular Value Decomposition

Theorem

Any mxn matrix can be put into the factored form **PDQ**, *where* **P** *and* **Q** *are unitary and* **D** *is diagonal. [2]*

A complex matrix U is unitary if

$$UU^H = U^H U = I$$

Singular Value Decomposition

Given matrix A, the SVD is:

$$A = PDQ$$

P: A matrix of the left singular vectors of A
D: A diagonal matrix with the singular values on its diagonal
Q: A matrix of the right singular vectors of A
Written in matrix form:

$$A = PDQ$$

$$\begin{bmatrix} . & . \\ . & . \\ . & . \end{bmatrix} = \begin{bmatrix} . & . & . \\ . & . & . \\ . & . & . \end{bmatrix} * \begin{bmatrix} a & . \\ . & b \\ . & . \end{bmatrix} * \begin{bmatrix} . & . \\ . & . \end{bmatrix}$$

Singular Value Decomposition

Singular Vectors: The left and right SVs combined span all of matrix A

- Left Singular Vectors: Spans the column space of A. The normalized eigenvectors of AA^T
- Right Singular Vectors: Spans the row space of A. The normalized eigenvectors of A^TA

Singular values: The square roots of the eigenvalues of $AA^T$and A^TA

The largest singular vectors explain the most variance of matrix A, and the singular values associated with each singular vector tell us how much variance is explained.

Singular Value Decomposition

Why are we referring to AA^T and A^TA instead of just matrix A?

- A is not necessarily square, and may not have positive eigenvalues that exist
- The matrices AA^T and A^TA are positive semidefinite - which means they are symmetric, have positive eigenvalues, and eigenvectors are pairwise orthogonal
- Still useful to find the eigenvectors and eigenvalues of AA^T and A^TA

SVD in Relation to PCA

PCA is built on SVD. Eigenvectors, known as components, are retained based on the amount of variance that they explain (quantified by eigenvalues).

PCA Overview

The four goals of PCA [1]

- Extract only the most important information from the data set
- Compress/ reduce the size of the data set via keeping only the important information
- Simplify the description of the data set
- Analyze the structure of (and relationships between) the variables

PCA Overview

- PCA uses the vectors in the original data table and finds a set of new, orthogonal vectors, called **Principal Components**
- Original vectors can be expressed as linear combinations of the new principal components.
- The primary principal component captures the largest variance

Change of Basis and Dimension Reduction

The original basis of the dataset before PCA is done has one variable per axis = confusing

After PCA, the basis is usually the first principal components. Therefore, we can create plots such as biplots which simply have a principal component on the x and y axis = less confusing

Applying PCA to Iraqui Migration Data

PCA was performed using R on a data set. This data set holds the results of a survey sourced from the International Organization for Migration. The survey was conducted in Iraq asking internally displaced persons about their access to certain needs, such as distance from clinics, access to clean water, etc.

- 206 variables
- 3,718 observations
- Collected April- June 2022

Data Cleaning

- Identify quantitative variables
- Remove commas from values with more than 3 digits
- Rename variables
- Create smaller dataframes with desired variables
- Addressed missing data... replaced missing with 0 for Residence good/ bad

Data Cleaning

Noticed that the number of families and number of individuals were actually just scaled by 6...

Step 1: Calculate Correlation Matrix

Correlation Matrix: In order to standardize data, the original matrix, A, is centered so that the mean of each column equals 0, and A will be normalized by dividing each variable by the norm.

- Therefore, values will range from 0 to $+/-1$.
- The diagonals will be 1 and this is a symmetric matrix. [1]

The code used in R to do this is shown below:

```
#Create a correlation matrix from the original (cleaned) data set A
cor1<-cor(originalData)

#Print this correlation matrix
cor1

```

Step 2: Use SVD to find Eigenvectors and Eigenvalues of Correlation Matrix

SVD is used to factor the correlation matrix into a form that shows the eigenvectors and eigenvalues of the correlation matrix.

Step 3: Get First Principal Component from Largest Eigenvalue

The largest eigenvalue tells us the eigenvector associated with capturing the most variance. This eigenvector becomes the first Principal Component.

R Code

Running PCA in R on the correlation matrix is shown below.

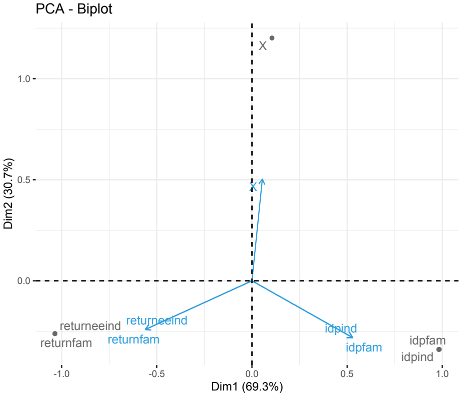
```r
#Run PCA on the correlation matrix
data.pca<-princomp(cor_matrix)

#Print results to show each component
data.pca
```

Results

Biplot: Shows the loadings on a graph to help visualize relationships between variables and interpret the principal components.

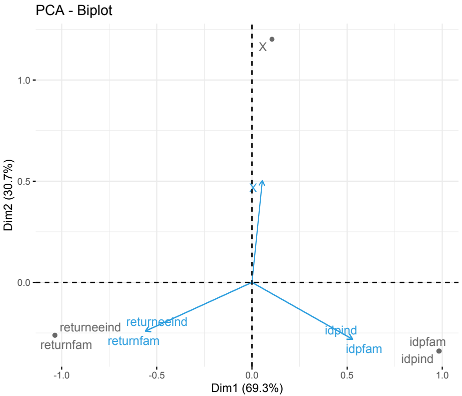
- **Loadings:** The length and the direction of the vectors. The loadings can tell us how much each variable contributes to a certain principal component

Initial Results



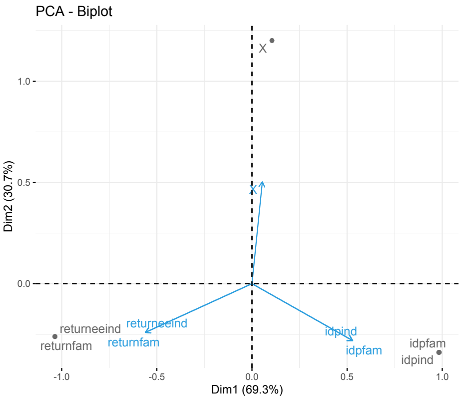
- X-axis: PC1, y-axis: PC2
- vector direction: which PC that variable contributes more to
- vector length: how much the variable contributes to a PC

Initial Results



Oops! Noticed that I accidentally included the index as a variable for my first run of PCA on a mini dataframe of 4 variables.

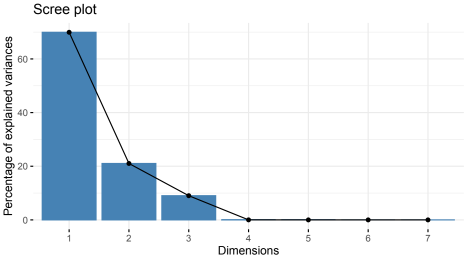
Initial Results



- the variable for the index, X, contributes almost entirely to PC2
- the number of individuals and families are slightly negatively correlated to PC2

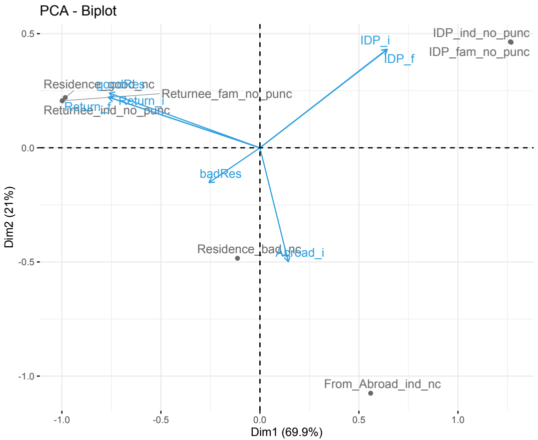
Results

Scree plot



A scree plot graphs the explained variance versus the principal component number.

Biplot


PCA - Biplot
IDP_i
IDP_f
IDP_ind_no_punc
IDP_fam_no_punc
Residence_good_nc
goodRes
Returnee_fam_no_punc
Return_f
Return_i
Returnee_ind_no_punc
badRes
Residence_bad_nc
Abroad_i
From_Abroad_ind_nc
Dim2 (21%)
Dim1 (69.9%)
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0

Takeaways

- Variables for number of individuals and families are multiples of each other
- Residence good and residence bad load on opposite direction vectors
- PCA uses linear algebra to help us understand a dataset and make it less complicated
- Data cleaning is very involved

Principal Components Analysis: Using Math to Extract Underlying Structures of Iraqi Migration Survey Data

Jackie Lindstrom

References

[1] Herve Abdi and Lynne J. Williams. "Principal Component Analysis". In: *WIREs Comp Stat* 2.4 (2010), pp. 433–459. DOI: 10.1002/wics.101.

[2] Ward Cheney and David Kincaid. *Linear Algebra Theory and Applications*. Jones and Bartlett Publishers, 2009. ISBN: 9780763750206.

Acknowledgements

- International Organization for Migration and H. Tran: collecting and sharing their data with me
- Dr. Justice for patience, support, and tea
- PLU Math Department

Positive Semi-Definite Matrices

A **positive semi-definite matrix** P can be obtained as the product of a matrix X and its transpose X^T.

$$P = XX^T \tag{5}$$

The properties of a positive semi-definite matrix are:

- Symmetric (perhaps define at beginning?)
- Positive eigenvalues
- Pairwise orthogonal eigenvectors

Positive Semi-Definite Matrices

In order to connect this to eigenvalues and eigenvectors, we recall U, the matrix where each column is an eigenvector of A, and Λ, the matrix that stores the eigenvalues of A. We can write P as

$$P = U\Lambda U^{-1} \tag{6}$$

Finding Eigenvalues

Therefore, finding the eigenvalues and eigenvectors for a matrix A involves finding values for $\vec{v}$ and λ that satisfy this equation. We can rewrite Equation 3 as

$$A\vec{v} = (\lambda I)\vec{v} \tag{7}$$

where I is the identity matrix with 1's down the diagonal. Now we have matrix multiplication on both sides of our equation.

Using algebra, we can rearrange equation 4 to give us:

$$(A - \lambda I)\vec{v} = 0 \tag{8}$$

This notation is helpful in showing that the eigenvectors, $\vec{v}$, remain on their span for some linear transformation represented by matrix A.

Unitary Definition

Definition

A real matrix U is orthogonal if

$$UU^T = U^T U = I$$

A complex matrix U is unitary if

$$UU^H = U^H U = I$$

[2]