Dimension Reduction (DR), Principal Component Analysis (PCA)

Ahmed Eleish

Data Analytics ITWS/MGMT 4600/6600 – CSCI/BCBP 4600

March 10th 2026

Tetherless World Constellation
Rensselaer Polytechnic Institute

Definitions

- **Dimensions of a dataset:** While a tabular dataset has two dimensions (or axes), which are its rows (vertical) and its columns (horizontal), the term *dimension,* when applied to a dataset most commonly refers to its columns, also called variables, attributes, or features. The two axes are often referenced as n x d, where n denotes the number of rows and d (sometimes p) denotes the number of dimensions,
  - These are analogous to the axes of a matrix (usually m x n).

- **Dimensionality Reduction (of a dataset):** the process of reducing the number of features of a dataset through some transformation that preserves the patterns or structure in the data.

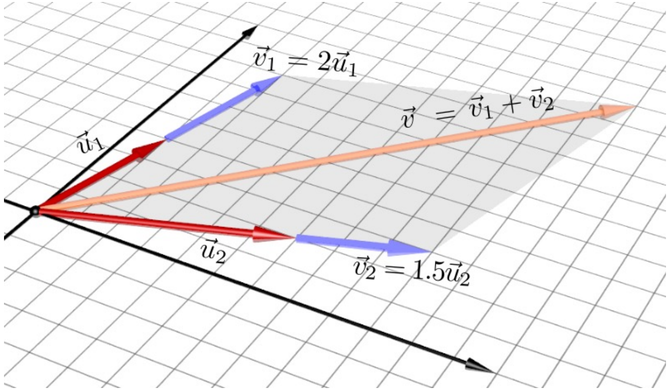
Definitions

- **Linear combination:** a mathematical expression where a set of terms of a vector are multiplied by a set of scalar constants and the results summed.

e.g.

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + a_3\mathbf{v}_3 + \cdots + a_n\mathbf{v}_n.$$

Where v_1 … v_n are vectors and a_1 … a_n are scalars



v is the linear combination of vectors u_1 and u_2 such that v = 2 * u_1 + 1.5 * u_2

https://en.wikipedia.org/wiki/Linear_combination

Variance/Covariance

- Variance and Covariance are measures of the “spread” of a set of points around their center of mass (mean).

- Variance is measure of the deviation from the mean for points in one dimension.

- Covariance is a measure of how much multiple dimensions vary from their means with respect to each other.

- Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions e.g. number of hours studied & marks obtained.

- The covariance between one dimension and itself is the variance

Variance/Covariance

Example:

- $VAR(X) = \frac{(12-36) + (20-36) + \ldots + (22-36)}{6}$

  = 862.333

- $COV(X,Y) = \frac{(12-36)*(1.7-4) + \ldots + (35-36)*(2.5-4)}{6}$

  = 77.283

X	Y
12	1.7
20	3.4
54	6.8
14	0.9
94	8.5
35	4.2
20	2.5

$$VAR(X) = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

$$COV(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Dimensionality Reduction (DR)

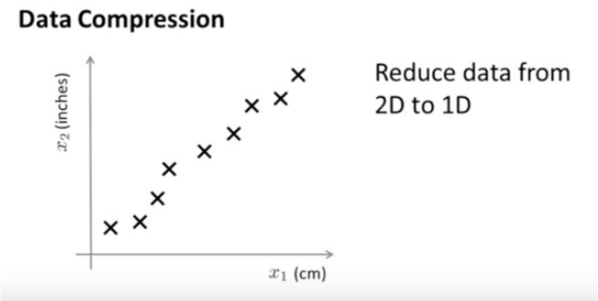
- There are multiple reasons that you want to do Dimensionality Reduction: one is to do compress data.

- Data compression not only saves memory space, it also speeds up learning algorithms.

Dimensionality Reduction (DR)

X1: distance measured in cm
X2: distance measured in inches

We want to reduce the data to one dimension



Length in cm is rounded off to the nearest cm and length in inches is rounded off to the nearest inch, that is why those examples do not perfectly lie on a straight line.

Image source; ML course Stanford University

Dimensionality Reduction (DR)

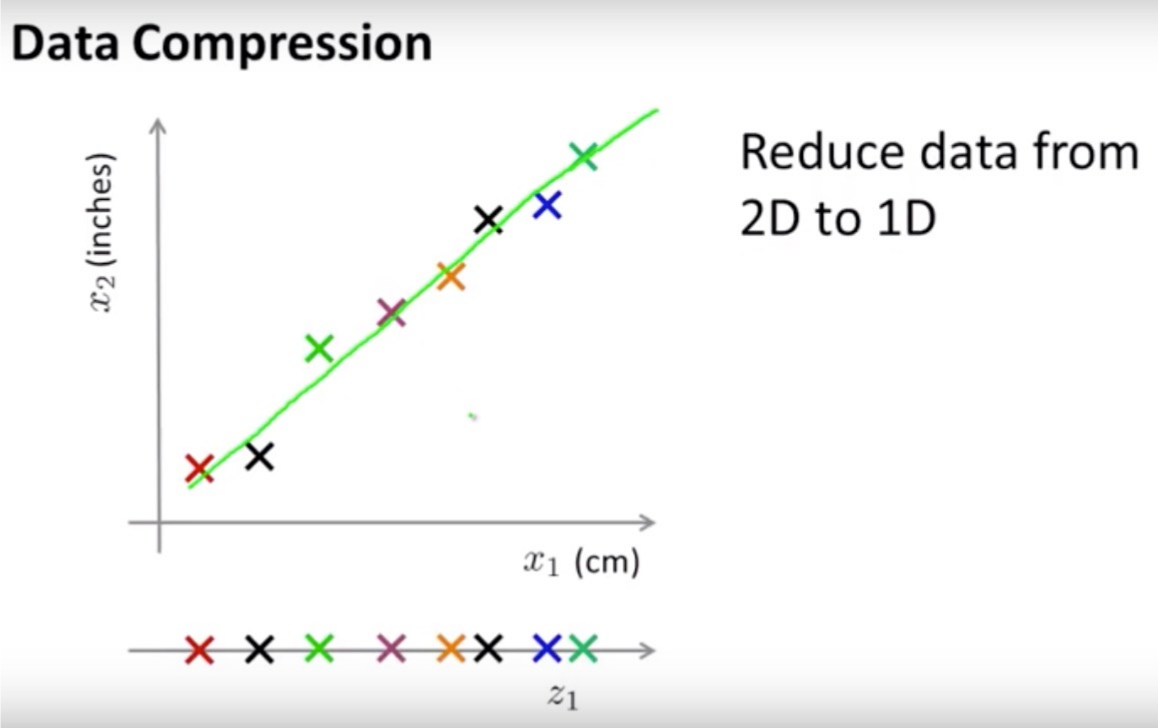


Image source; ML course Stanford University

Dimensionality Reduction (DR)

- If we allow ourselves to approximate the original dataset by projecting all the original examples onto the green line, then we need only one number to specify a point on the line.
- This way, we have reduced the problem from 2D to 1D.
- For this example this may not be a big deal, but if you have a dataset with large number of features with redundant information, it will take too much memory space and take more time to do the computations.
- In that case it's better to reduce the redundancy.

Dimensionality Reduction (DR)

- If a dataset contains hundreds of features, it is difficult to keep track of all those features, and there is redundancy.

-> DR is used in feature selection, reduction

- Why?
    - **Curse of dimensionality** – challenges of learning from (very) high-dimensional data, or datasets with a large number of dimensions relative to the number of observations*.

* These considerations are generalizations and are more narrowly defined per domain/dataset/analysis

Dimensionality Reduction Methods

- Linear:
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Factor Analysis (FA)
- Non-linear:
  - Uniform Manifold Approximation and Projection (UMAP)
  - Autoencoders (Neural Networks)
  - Kernel PCA

Principal Component Analysis (PCA)

Principal Component Analysis

- Principal Component Analysis (PCA) is an unsupervised learning technique.
- PCA is a popular approach for deriving a low-dimensional set of features from a large set of variables.
- The core concept is that a large part of the variation in a dataset can be explained using fewer variables called “Principal Components” that are linear combinations of the original variables.

Dimensionality Reduction with PCA

PCA is useful in many different scenarios, including:

- **Data exploration:** PCA can help you to visualize high- dimensional data in a lower dimensional space.
- **Data compression:** PCA can reduce the number of variables in a dataset, which can make it easier to work with.
- **Feature selection:** PCA can help to identify the most important variables in a dataset.
- **Data pre-processing:** PCA can be used to remove noise from a dataset and to standardize variables so that they have a similar scale.
- **Downstream Machine learning:** PCA can be used as a pre-processing step before applying machine learning algorithms to a dataset, to improve their performance and reduce overfitting.

Principal Component Analysis

- Suppose that we wish to visualize *n* observations with measurements on a set of *p* features, *X1, X2, X3, ..., Xp* as a part of exploratory data analysis.

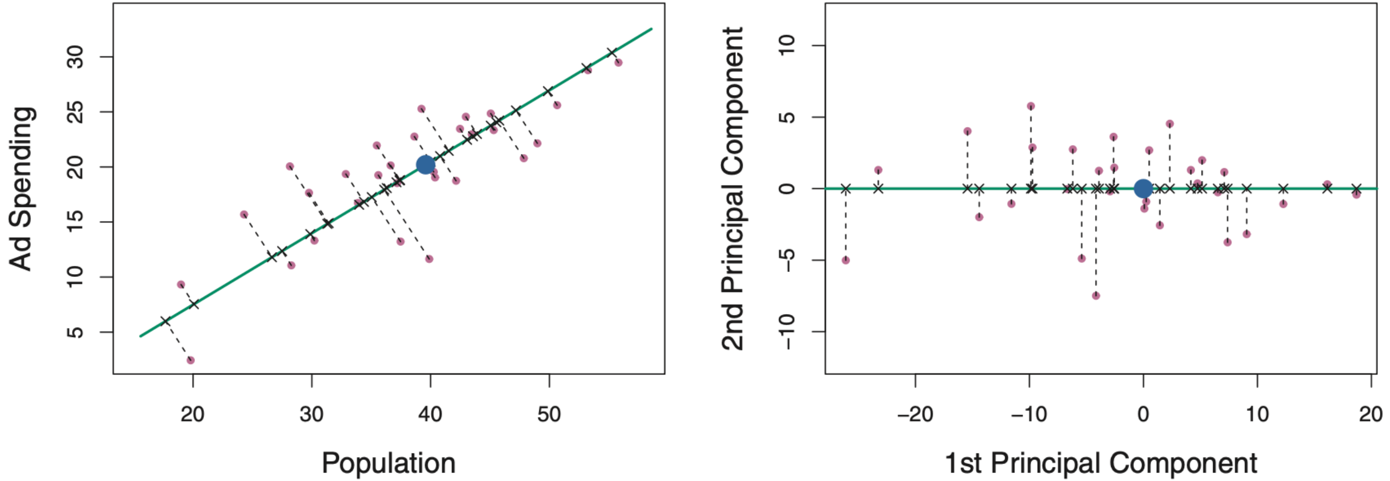
- We could do this by examining two-dimensional scatterplots of the data, which contain measurements on two of the features for *n* observations. There are C_2^p of such scatterplots, for example with *p=10*, there are 45 plots!

$$C_2^p = \frac{p(p-1)}{2}$$

- If p is large, then it will certainly not be possible to look at all of them. Moreover, most likely, many of them will not be informative since they each contain just a small fraction of the total information present in the dataset.

Principal Component Analysis

- The dataset is projected onto the Principal Components for visualization or further analysis:



Principal Components

- The principal components of a dataset are calculated by taking **linear combinations of the original variables** in such a way that *each component is orthogonal (uncorrelated) to the others*.

- Eliminating less significant principal components allows us to represent the data in a lower-dimensional space, which is easier to understand and analyze.

Finding the Principal Components

The first principal component of a set of features X1, X2, . . . , Xp is the normalized linear combination of the features that has the largest **variance**.

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p$$

By normalized, we mean that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$ We refer to the elements $\varphi_{11},\ldots,\varphi_{p1}$ as the loadings of the first principal component; together, the loadings make up the principal component loading vector, $\varphi_1 = (\varphi_{11}, \varphi_{21}, \ldots, \varphi_{p1})^T$.

Once the loading vectors for all principal components are obtained, the dataset can be transformed (projected) to obtain new values according to the new axes (PCs).

$$Z = XV$$

where *Z* is the transformed matrix, *X* is the original matrix, *V* is the matrix of loading vectors ($\varphi_1, \ldots \varphi_p$)

Finding the Principal Components

Steps:

1) Center dataset: subtract column means from each row such that the mean of each variable is 0.

$$X_c = X - \bar{X}$$

Optionally standardize variables by dividing each variable by its standard deviation: $X_s = \frac{X_c}{SD(x)}$

2) Find covariance matrix: $C = \frac{1}{N-1} X_s * X_s$

3) Factorize the covariance matrix using eigen-decomposition to find the its eigenvectors and eigenvalues.

4) Rearrange eigenvectors descendingly by eigenvalues. The rearranged eigenvectors are the loadings of the principal components of the original (centered and scaled) data matrix *Xs*.

Eigen-decomposition

- It is the factorization of a matrix into a product of simpler matrices, analogous to factoring an integer into 2 integers, e.g. 12 = 3 * 4

$$C = V\Lambda V^T$$

C is a square matrix (e.g. a covariance matrix), ***V*** is a square matrix of *eigenvectors* and **Λ** is a diagonal matrix of *eigenvalues* ($\lambda_1, \ldots, \lambda_n$)

- An eigenvector is a vector that when multiplied by a square matrix, it is scaled it by a scalar value called an eigenvalue

$$Cv = \lambda v$$

- The vectors in the matrix V (eigenvectors of the covariance matrix) are the loadings of the principal components of the original data matrix and the eigenvalues are the variances of the principal components.

In-Class Work examples

- PCA on Iris dataset.

  https://rpi.box.com/s/xut9h86qwsehlry41dq3eagu3v5p5o1p

PCA on Boston dataset

```
install.packages('MASS')
boston.df <- Boston
```

Do PCA!

Thanks!