CS 5683: Algorithms & Methods for Big Data Analytics

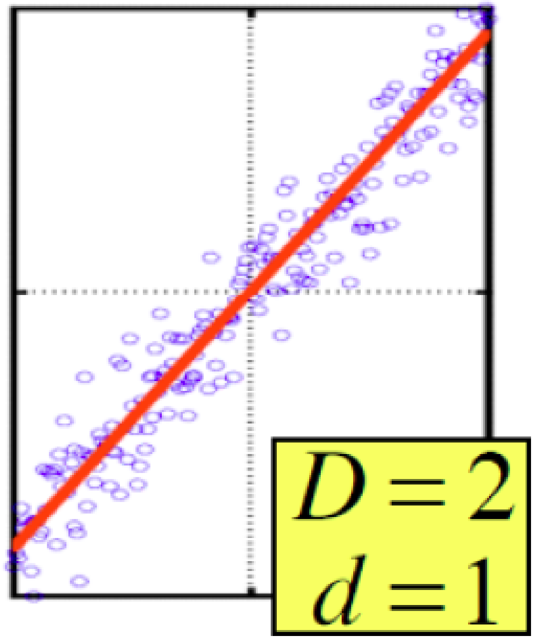
Dimensionality Reduction – 1
PCA

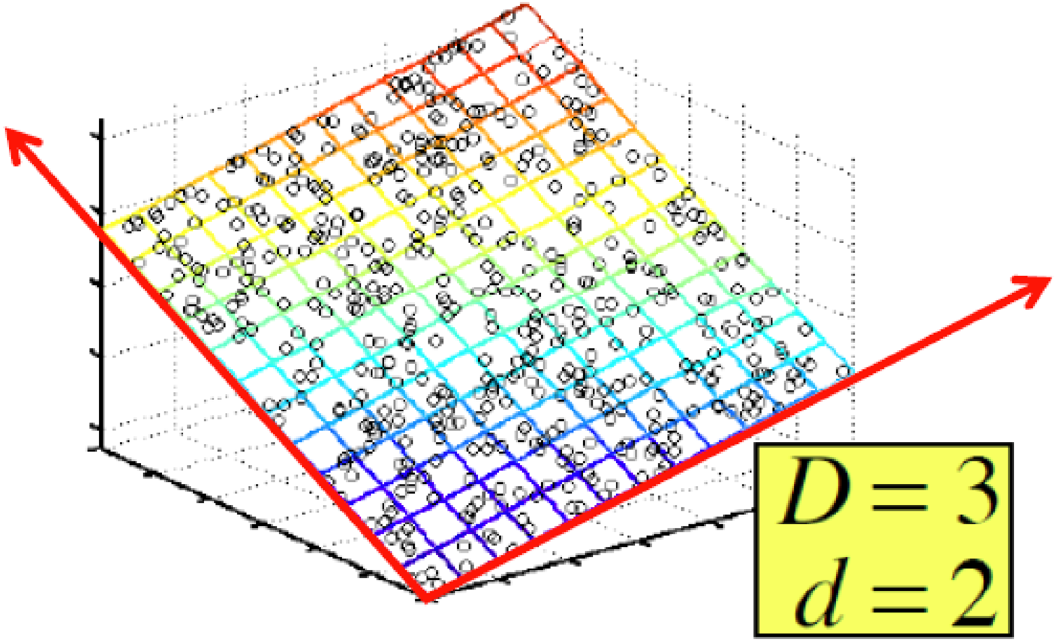
Arunkumar Bagavathi
Department of Computer Science
Oklahoma State university

Topics Overview

High. Dim. Data	Text Data	Graph Data	Others
Data Features	Clustering	PageRank	Data Streams Mining
Dimensionality Reduction	Non-linear Dim. Reduction	ML for Graphs	Intro. to Apache Spark
Application Rec. Systems	*Application* IR	Community Detection	

Dimensionality Reduction





- **Assumption:** Data lies on or near a low *d*-dimensional subspace
- **Axes of this subspace are effective representation of the data**

Dimensionality Reduction

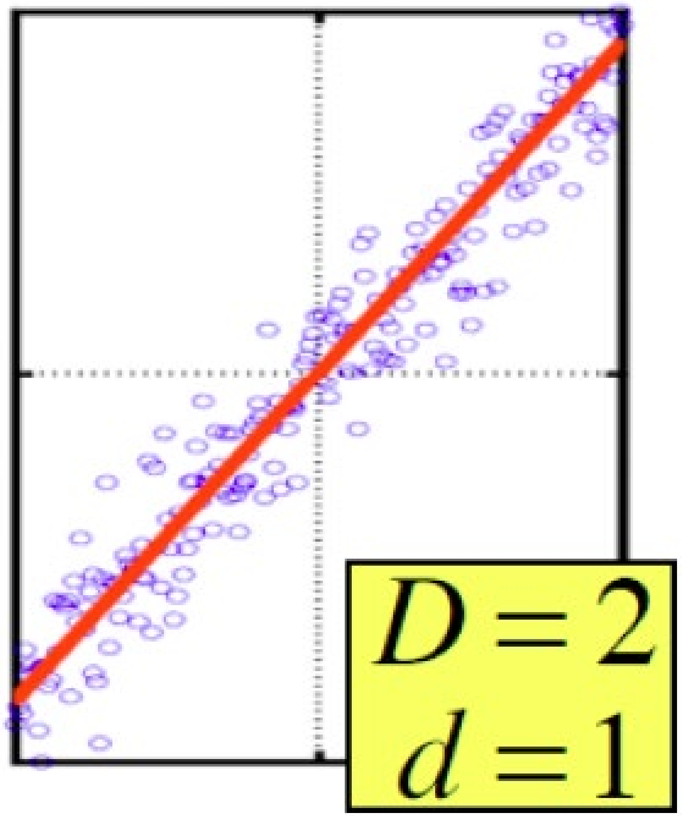
- **Compress / reduce dimensionality:**
  - 10^6 rows; 10^3 columns; no updates
  - Random access to any cell(s); **small error: OK**

day / customer	We 7/10/96	Th 7/11/96	Fr 7/12/96	Sa 7/13/96	Su 7/14/96
ABC Inc.	1	1	1	0	0
DEF Ltd.	2	2	2	0	0
GHI Inc.	1	1	1	0	0
KLM Co.	5	5	5	0	0
Smith	0	0	0	2	2
Johnson	0	0	0	3	3
Thompson	0	0	0	1	1

The above matrix is really "2-dimensional." All rows can be reconstructed by scaling [1 1 1 0 0] or [0 0 0 1 1]

Dimensionality Reduction

- **Goal of dimensionality reduction: to discover the axis of data!**



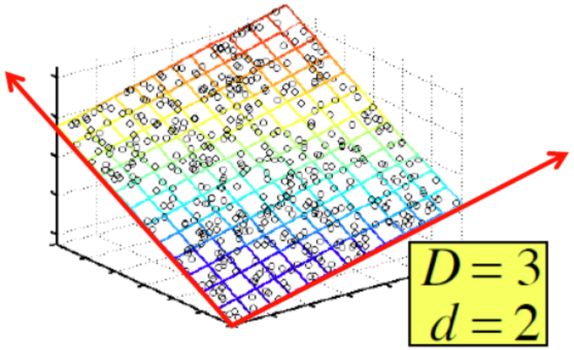
Rather than representing every point with 2 coordinates we represent each point with 1 coordinate (corresponding to the position of the point on the red line).

By doing this we incur a bit of **error** as the points do not exactly lie on the line

Why Reduce Dimensions

Why reduce dimensions?

- **Discover hidden correlations/topics**
  - Words that occur commonly together
- **Remove redundant and noisy features**
  - Not all words are useful
- **Interpretation and visualization**
- **Easier storage and processing of the data**



Linear Algebra Throwback

1. **$M_{m \times n}$ – matrix with *m rows* and *n columns*:** $M = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$
   - **Diagonal matrix** – matrix with 0's everywhere except the diagonal: $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$
   - **Symmetric matrix** - $\boldsymbol{M = M^T}$ **(i.e) $M_{i,j} = M_{j,i}$** (M should be a square matrix)
   - **Identity matrix (I)** – diagonal matrix with only 1's in the diagonal: $\mathrm{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
   - **Orthogonal matrix** – matrix is orthogonal if $\boldsymbol{MM^T = M^TM = I}$
     - If M is orthogonal, then M^T is also orthogonal
     - All column (or row) vectors in an orthogonal matrix have unit length – sum of squares of its elements =1
     - Dot product of two column (or row) vectors = 0

Linear Algebra Throwback

2. Eigen vector – if a matrix is multiplied by a vector (**x**) and the vector x gets linearly transformed (stretched, without changing the direction), then the vector x is called **Eigen vector**

3. Eigen value – quantity (λ) at which the vector x is transformed: $\mathbf{Mx} = \lambda \mathbf{x}$

4. Properties:

- All Eigen vectors are unit vectors
- Determinant of a matrix = the product of its Eigen values
- Eigen vectors in the Eigen matrix are orthogonal (perpendicular to each other)

Linear Algebra Throwback

Solving for Eigen values and Eigen vectors of a square matrix $M = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$

$\mathbf{Mx} = \lambda\mathbf{x} \Rightarrow (\mathbf{M} - \lambda \boldsymbol{I})\mathbf{x} = \mathbf{0}$ (this condition holds **iff** $|\mathbf{M} - \lambda \boldsymbol{I}| = \mathbf{0}$)

$M - \lambda I = \begin{bmatrix} 3-\lambda & 2 \\ 2 & 6-\lambda \end{bmatrix} \Rightarrow |M - \lambda I| = (3-\lambda)(6-\lambda) - 4 = 0$

$\lambda = 7$ (largest Eigen value – **principal Eigen value**) and $\lambda = 2$

Solve for first Eigen vector $\begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = 7\begin{bmatrix} x \\ y \end{bmatrix}$

$$\mathbf{3x + 2y = 7x}$$
$$\mathbf{2x + 6y = 7y}$$
$$\mathbf{y = 2x}$$

Linear Algebra Throwback

$$\boldsymbol{y = 2x}$$

Possible Eigen vector can be $\begin{bmatrix} \mathbf{1} \\ \mathbf{2} \end{bmatrix}$ (**Remember, Eigen vectors are unit vectors!**)

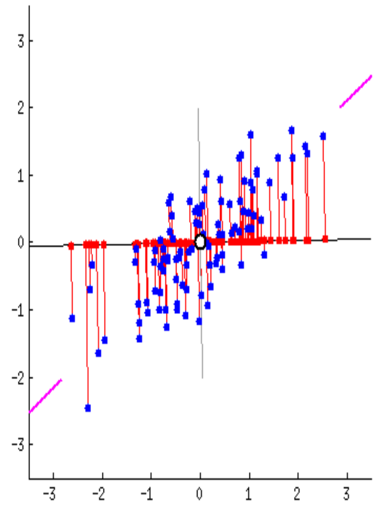
Thus, Eigen vector is $\begin{bmatrix} \mathbf{1/\sqrt{5}} \\ \mathbf{2/\sqrt{5}} \end{bmatrix}$ (**Principal Eigen vector**)

Similarly for the other Eigen value λ = 2, the Eigen vector is $\begin{bmatrix} \mathbf{2/\sqrt{5}} \\ \mathbf{-1/\sqrt{5}} \end{bmatrix}$

Matrix of Eigen vectors $\mathrm{E} = \begin{bmatrix} \mathbf{1/\sqrt{5}} & \mathbf{2/\sqrt{5}} \\ \mathbf{2/\sqrt{5}} & \mathbf{-1/\sqrt{5}} \end{bmatrix}$ ➔ $\mathrm{EE^T} = \mathrm{E^TE} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}$

Principal Component Analysis (PCA)

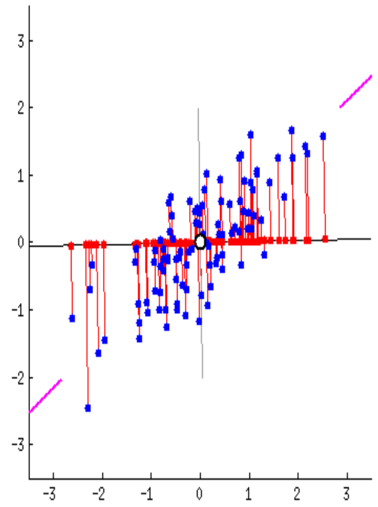
- A data mining technique that identifies the projection of the high-dimensional data onto which the data tuples align the best
- **In other words:** Find Eigen vectors (*components*) of the original data – the *Principal Eigen vector* represents an axis where most data points reside (*high variance*)



PCA Step - 1

- We cannot apply PCA to the original data ***M*** (since the variables of this matrix can be *different scales*)
- ***Example:*** Variable of scale 1 – 10 and Variable of scale 20 – 200
- **Standardization:** We standardize the data to keep all variables in same scale using Z-Score

$$x_i = \frac{x_i - \bar{x}}{\sigma_x}$$



- Z-Score interprets 'p' standard deviations from mean
  - Positive – above the mean
  - 0 – mean
  - Negative – below the mean

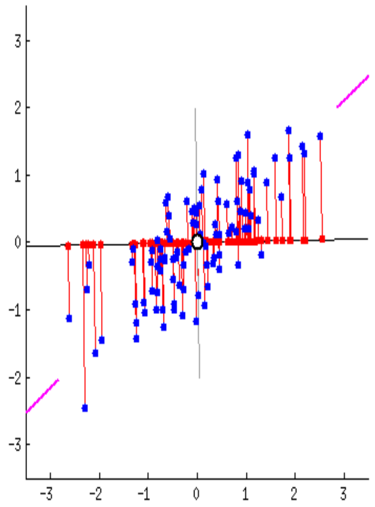
Let the standardized data be **M'**

PCA Step - 2

- We cannot apply PCA to the original data ***M'*** (since this matrix can be ***n x m| n***≠m)
- So, we apply it on the corresponding correlation matrix $\mathbf{M}''$
- **Pearson Correlation –** Measures linear relationships and dependencies between two features

$$M''_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{n}(x_i - \bar{x})^2 \sum_{i=0}^{n}(y_i - \bar{y})^2}}$$

- Pearson correlation projects the data into $m \times m$ space



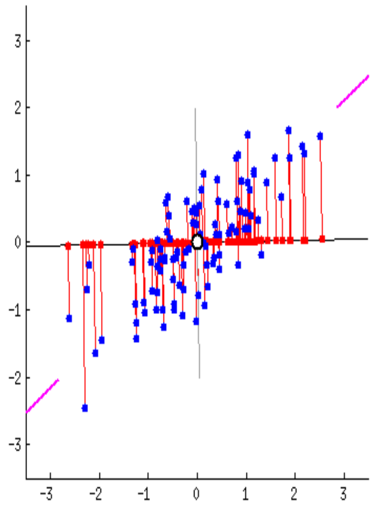
Students task: Investigate about the Covariance matrix

PCA Step - 3

- Eigen vectors from the correlation matrix M'' can be considered as a rotation of a high-dimensional space with relation to eigen values

- **PCA Idea:** Data points along the principal Eigen vector are most spread out (variance is maximized)

- **Eigen matrix (E):** 'k' principal eigen vectors organized by their magnitude



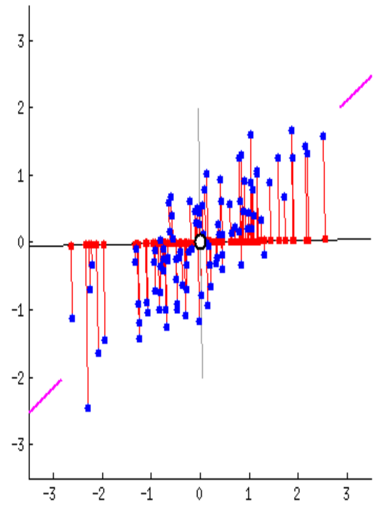
Students task: Choose optimal 'k' in Assignment-1

PCA Step - 4

- Project the original data in a low-dimensional coordinate space by:

$$\widehat{M} = M'.E$$

- $\widehat{M} \rightarrow$ the first axis corresponds to the largest eigen value, the second axis corresponds to the second largest eigen values, and so on

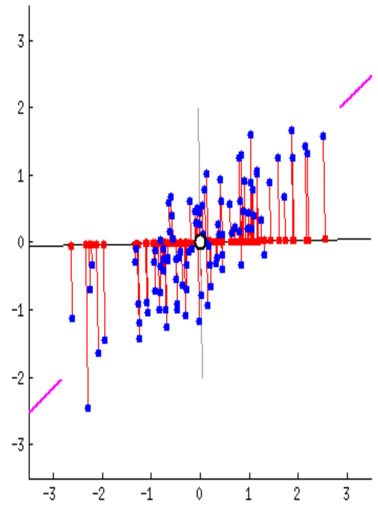


PCA Step - 5

- Reconstruct the original data in a low-dimensional coordinate space by:

$$\widehat{F} = \widehat{M}.E^T$$

- The reconstructed data gives an approximation of the original data from the low-dimensional space

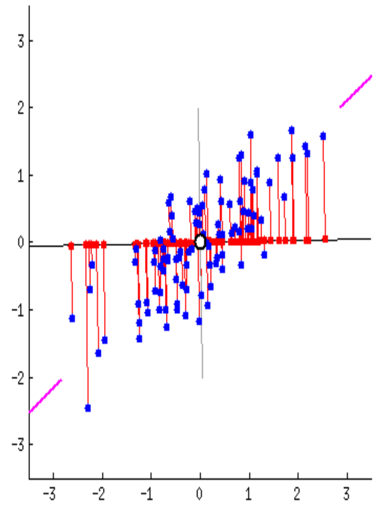


Students task: Do not forget the de-standardization for full re-construction

PCA Step - 6

- **Reconstruct loss:** Evaluate the reconstructed data with Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})$$



Questions???

Acknowledgements

Some of the slides of this lecture are inspired from the Mining Massive Datasets course: http://www.mmds.org/