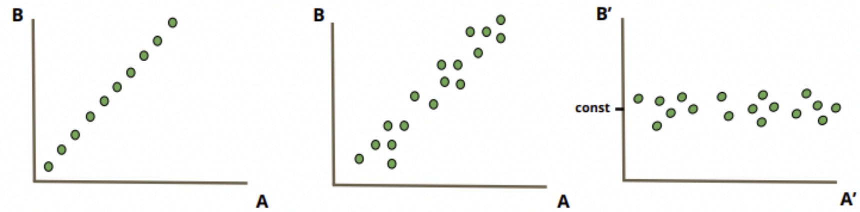
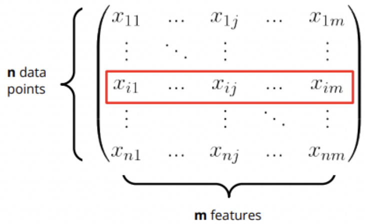
CAS CS 506
Lec 09

Singular Value Decomposition

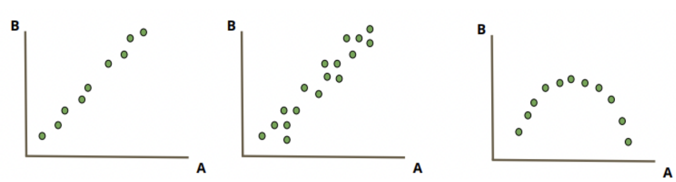
1. Characteristics of a Dataset to Look for
    a. 

    b.
$$\text{n data points}\left\{\underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m}\\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm}\end{pmatrix}}_{\text{m features}}\right.$$
2. Goal
    a. Examine this matrix and uncover its linear algebraic properties to
        i. Approximate A with a smaller matrix B that is easier to store but contains similar information as A
$$\text{n data points}\left\{\underbrace{\begin{array}{c} \begin{array}{ccccc} \mathbf{A} & \dots & \mathbf{J} & \dots & \mathbf{M} \end{array} \\ \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m}\\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm}\end{pmatrix}\end{array}}_{\text{m features}}\right. \quad \underbrace{\begin{array}{c} \begin{array}{ccccc} \mathbf{A'} & \dots & \mathbf{J'} & \dots & \mathbf{M'} \end{array} \\ \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m}\\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm}\end{pmatrix}\end{array}}_{\text{m features}}$$
        ii. Dimensionality Reduction / Feature Extraction
$$\text{n data points}\left\{\underbrace{\begin{array}{c} \begin{array}{ccccc} \mathbf{A} & \dots & \mathbf{J} & \dots & \mathbf{M} \end{array} \\ \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m}\\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm}\end{pmatrix}\end{array}}_{\text{m features}}\right. \quad \underbrace{\begin{array}{c} \begin{array}{ccc} \mathbf{A''} & \dots & \mathbf{J''} \end{array} \\ \begin{pmatrix} x_{11} & \cdots & x_{1j}\\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj}\end{pmatrix}\end{array}}_{\text{j features}}$$

iii. Anomaly Detection & Denoising



3. Linear Algebra Review
   a. Definition: The vectors in a set V = {v_1 , …, v_n} are linearly independent if
      $a_1 * v_1 + \ldots + a_n * v_n = o$
      can only be satisfied by $a_i = 0$
      i. Notice: this means no vector in that set can be expressed as a linear combination of other vectors in the set
   b. The determinant of a square matrix A is a scalar value that encodes properties about the linear mapping described by A
   c. n vectors {v_1 , …, v_n} in an n-dimensional space are linearly independent iff the matrix A:
      A = [v_1, …, v_n] (n x n)
      has non-zero determinant
   d. The rank of a matrix A is the dimension of the vector space spanned by its column space. This is equivalent to the maximal number of linearly independent columns / rows of A
      i. A matrix A is full-rank iff rank(A) = min(m, n)
4. Matrix Factorization
   a. Any matrix A of rank k can be factored as
      A = UV
      where
      U is n * k
      V is k * m
   b. To store an n * m matrix A requires storing m * n values
   c. However, if the rank of the matrix of A is k, since A can be factored as
      A = UV
      which requires storing k(m + n) values
5. In Practice
   a. Most datasets are full rank despite containing a lot of redundant / similar information…



   b. But we might be able to approximate the dataset with a lower rank one that contains similar information
6. Approximation
   a. Goal
      i. Approximate A with A^(k) (low-rank matrix) such that
         1. d(A, A^(k)) is small
         2. k is small compared to m & n
7. Frobenius Distance
   a. $$d_F(A, B) = \|A - B\|_F = \sqrt{\sum_{i,j}(a_{ij} - b_{ij})^2}$$
   b. I.e. the pairwise sum of squares difference in values of A and B
8. Approximation
   a. Definition:
   b. When k < rank(A), the rank-k approximation of A (in the least squares sense) is
   $$A^{(k)} = \underset{\{B|rank(B)=k\}}{\arg\min} \; d_F(A, B)$$
9. Matrix Factorization Improved
   a. Not only can we factorize a matrix A of rank k as A = UV. But we can factorize A using a process called Singular Value Decomposition (SVD) where
   A = U**Σ**V^T
10. Approximation
   a. Definition:
   The singular value Decomposition of a rank-r matrix A has the form
   A = U**Σ**V^T
   where
   U is n * r
   The columns of U are orthogonal & unit length (U^T * U = I)
   V is m * r
   The columns of V are orthogonal & unit length (V^T * V = I)
   b. where
   $$\Sigma = \begin{pmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_r \end{pmatrix}$$
   with $\boldsymbol{\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0}$

   $\boldsymbol{\sigma_i}$ is the square root of the eigenvalues of $\mathbf{A^TA}$ and are called **singular values**

c. Find A^(k) by decomposing A:

$$A = \begin{pmatrix} \boxed{U_1} & U_2 \end{pmatrix} \begin{pmatrix} \boxed{\Sigma_1} & \\ & \Sigma_2 \end{pmatrix} \begin{pmatrix} \boxed{V_1} & V_2 \end{pmatrix}$$

$A^{(k)} = U_1\Sigma_1 V_1^T$

Where

U_1 is **n x k**
Σ_1 is **k x k**
V_1 is **m x k**

d.

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
0	0	0	2	2
0	0	0	3	3
0	0	0	1	1

~

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

x

9.64	0
0	0

x

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

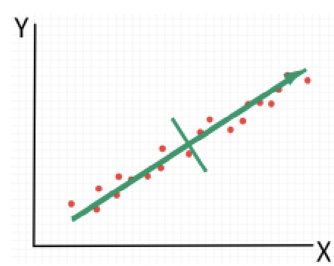
1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
0	0	0	2	2
0	0	0	3	3
0	0	0	1	1

~

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

e. The ith singular vector represents the direction of the ith most variance

$$\Sigma = \begin{pmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_r \end{pmatrix}$$



f. Singular Values express the importance / significance of a singular vector

g. Property:

$$d_F(A, A^{(k)})^2 = \sum_{i=k+1}^{r} \sigma_i^2$$

i. Note: the larger k is, the smaller the distance

h. To find the right k, you can:

i. Look at the singular value plot to find the elbow point

ii. Look at the residual error of choosing different k

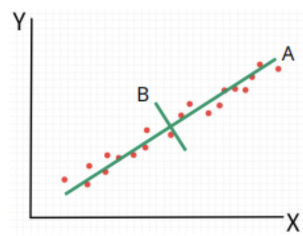
11. Related to Principal Component Analysis (PCA)

a. SVD and PCA are related

12. Dimensionality Reduction

a. Idea: project the data onto a subspace generated from a subset of singular vectors / principal components

b. Want to project onto the components that capture most of the variance / information in the data

c. 

i. Which principal component should we project on?

ii. A:

13. Anomaly Detection

a. Define O = A - A^(k)

b. The largest rows of O could be considered anomalies