THE CURSE OF DIMENSIONALITY

- Real data usually have thousands, or millions of dimensions
  - E.g., web documents, where the dimensionality is the vocabulary of words
  - Facebook graph, where the dimensionality is the number of users
- Huge number of dimensions causes problems
  - Data becomes very sparse, some algorithms become meaningless (e.g. density based clustering)
  - The complexity of several algorithms depends on the dimensionality and they become infeasible.

DIMENSIONALITY REDUCTION

- We will see other forms of dimensionality reduction
- LSH, and random projections reduce the dimension while preserving the distances

DATA IN THE FORM OF A MATRIX

- We are given n objects and d attributes describing the objects. Each object has d numeric values describing it.

- We will represent the data as a $n \times d$ real matrix A.
  - We can now use tools from linear algebra to process the data matrix

- Our goal is to produce a new $n \times k$ matrix B such that
  - It preserves as much of the information in the original matrix A as possible
  - It reveals something about the structure of the data in A

EXAMPLE: DOCUMENT MATRICES

d terms
(e.g., theorem, proof, etc.)

n
documents

$$A$$

A_{ij} = frequency of the **j**-th term in the **i**-th document

Find subsets of terms that bring documents together

EXAMPLE: RECOMMENDATION SYSTEMS

d movies

n customers

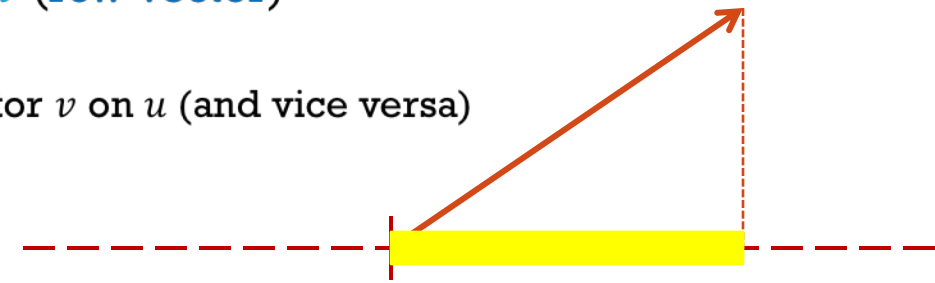
$$A$$

A_{ij} = rating of **j**-th product by the **i**-th customer

Find subsets of movies that capture the behavior or the customers

LINEAR ALGEBRA

- We assume that vectors are column vectors.
- We use v^T for the transpose of vector v (row vector)
- Dot product: $u^T v$ $(1\times n, n\times 1 \rightarrow 1\times 1)$
  - The dot product is the projection of vector v on u (and vice versa)



- $[1,2,3]\begin{bmatrix}4\\1\\2\end{bmatrix} = 12$
- $u^T v = \|v\|\|u\|\cos(u,v)$
  - If $||u|| = 1$ (unit vector) then $u^T v$ is the projection length of v on u
- $[-1,2,3]\begin{bmatrix}4\\-1\\2\end{bmatrix} = 0$ orthogonal vectors
  - Orthonormal vectors: two unit vectors that are orthogonal

MATRICES

- An n×m matrix A is a collection of n row vectors and m column vectors

$$A = \begin{bmatrix} | & | & | \\ a_1 & a_2 & a_3 \\ | & | & | \end{bmatrix} \qquad A = \begin{bmatrix} - & \alpha_1^T & - \\ - & \alpha_2^T & - \\ - & \alpha_3^T & - \end{bmatrix}$$

- Matrix-vector multiplication
  - Right multiplication Au: projection of u onto the row vectors of A, or projection of row vectors of A onto u.
  - Left-multiplication $u^T A$: projection of u onto the column vectors of A, or projection of column vectors of A onto u
- Example:

$$[1,2,3]\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = [1,2]$$

RANK

- Row space of A: The set of vectors that can be written as a linear combination of the rows of A
  - All vectors of the form $v = u^T A$
- Column space of A: The set of vectors that can be written as a linear combination of the columns of A
  - All vectors of the form $v = Au$.
- Rank of A: the number of linearly independent row (or column) vectors
  - These vectors define a basis for the row (or column) space of A

RANK-1 MATRICES

- In a rank-1 matrix, all columns (or rows) are multiples of the same column (or row) vector

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 4 & -2 \\ 3 & 6 & -3 \end{bmatrix}$$

- All rows are multiples of $r = [1,2,-1]$
- All columns are multiples of $c = [1,2,3]^T$
- External product: uv^T $(n\times 1\ ,\ 1\times m\ \rightarrow\ n\times m)$
  - The resulting $n\times m$ has rank 1: all rows (or columns) are linearly dependent
  - $A = rc^T$

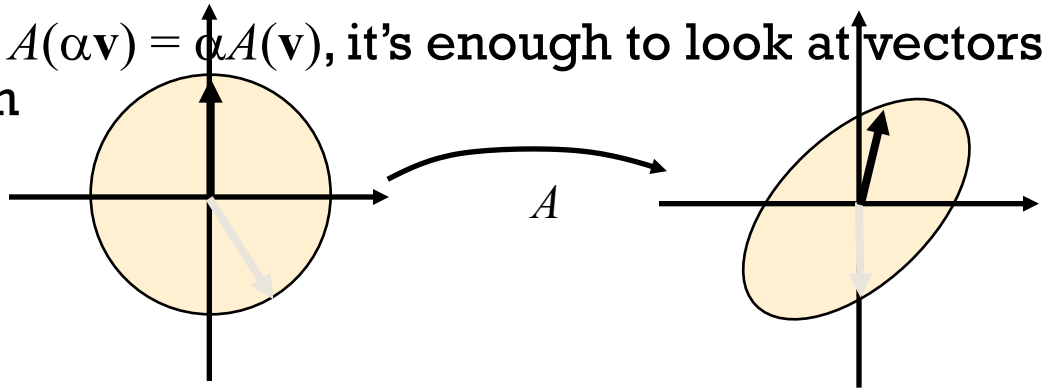
EIGENVECTORS

- (Right) Eigenvector of matrix **A**: a vector **v** such that $Av = \lambda v$
- λ: eigenvalue of eigenvector v

- A square matrix **A** of rank r, has r orthonormal eigenvectors $u_1, u_2, \ldots, u_r$ with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_r$.
- Eigenvectors define an orthonormal basis for the column space of **A**

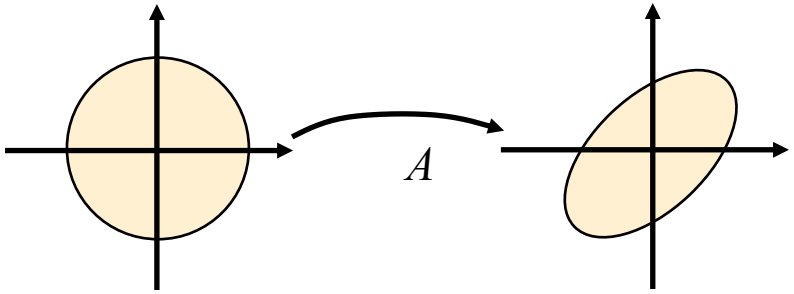
GEOMETRIC ANALYSIS OF LINEAR TRANSFORMATIONS

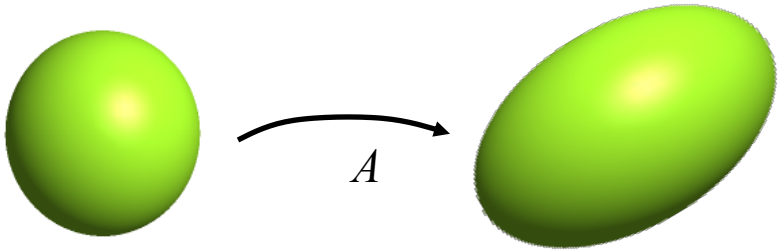
- We want to know what a linear transformation A does
- Need some simple and "comprehendible" representation of the matrix of A.
- Let's look what A does to some vectors
  - Since $A(\alpha \mathbf{v}) = \alpha A(\mathbf{v})$, it's enough to look at vectors v of unit length



THE GEOMETRY OF LINEAR TRANSFORMATIONS

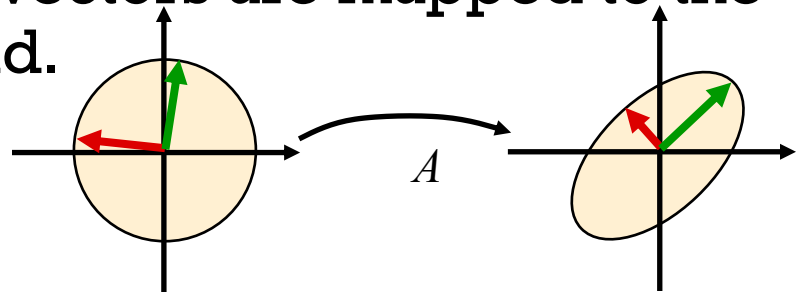
- A linear (non-singular) transform A always takes hyper-spheres to hyper-ellipses.

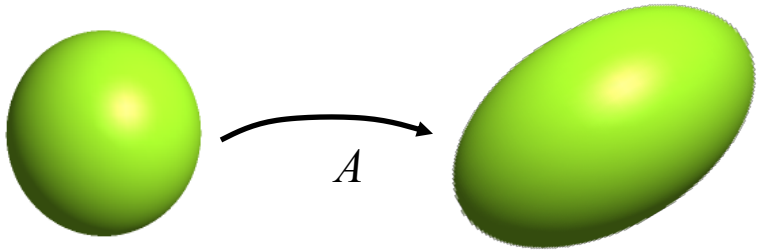




THE GEOMETRY OF LINEAR TRANSFORMATIONS

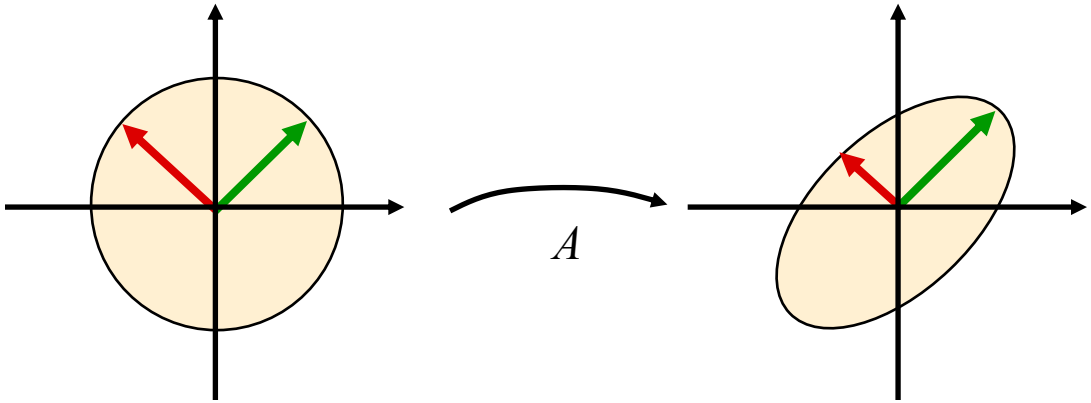
- Thus, one good way to understand what A does is to find which vectors are mapped to the "main axes" of the ellipsoid.





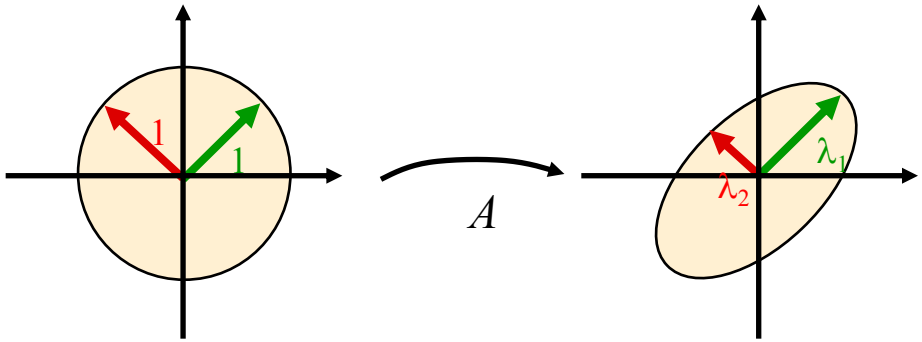
GEOMETRIC ANALYSIS OF LINEAR TRANSFORMATIONS

- If we are lucky: $A = V \Lambda V^T$, V orthogonal (true if A is symmetric)
- The eigenvectors of A are the axes of the ellipse



SYMMETRIC MATRIX: EIGEN DECOMPOSITION

- In this case A is just a scaling matrix. The eigen decomposition of A tells us which orthogonal axes it scales, and by how much:

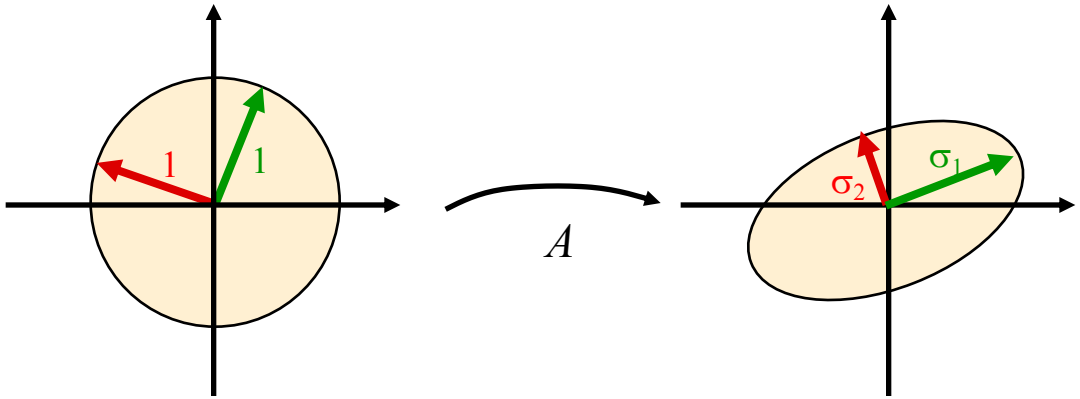


$$A = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{bmatrix}^T$$

$$\boxed{A\mathbf{v}_i = \lambda_i \mathbf{v}_i}$$

GENERAL LINEAR TRANSFORMATIONS: SVD

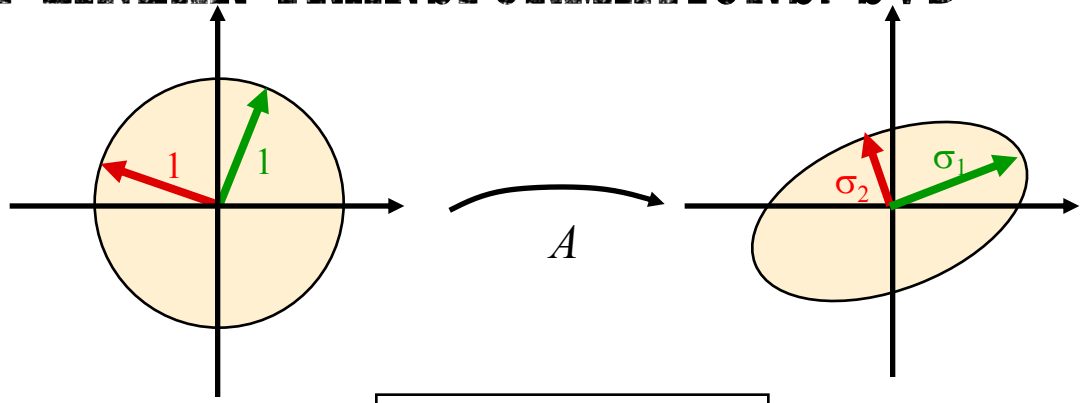
- In general A will also contain rotations, not just scales:



$$A = U \Sigma V^T$$

$$A = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_n \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{bmatrix}^T$$

GENERAL LINEAR TRANSFORMATIONS: SVD



$$AV = U\Sigma$$

$$A\begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n\end{bmatrix} = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_n\end{bmatrix}\begin{bmatrix}\sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n\end{bmatrix}$$

orthonormal *orthonormal*

$$A\mathbf{v}_i = \sigma_i\mathbf{u}_i, \quad \sigma_i \geq 0$$

SINGULAR VALUE DECOMPOSITION

- $$A = U \ \ \Sigma \ \ V^T = [u_1, u_2, \cdots, u_r] \begin{bmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_r \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_r^T \end{bmatrix}$$

$$[n \times m] = [n \times r] \ \ [r \times r] \ \ [r \times m]$$

r: rank of matrix A

- $\sigma_1, \geq \sigma_2 \geq \cdots \geq \sigma_r$: singular values of matrix A (also, the square roots of eigenvalues of AA^T and $A^T A$)

- $u_1, u_2, \ldots, u_r$: left singular vectors of A (also eigenvectors of AA^T)

- $v_1, v_2, \ldots, v_r$: right singular vectors of A (also, eigenvectors of $A^T A$)

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T$$

SYMMETRIC MATRICES

- Special case: A is symmetric positive definite matrix

$$A = \lambda_1 u_1 u_1^T + \lambda_2 u_2 u_2^T + \cdots + \lambda_r u_r u_r^T$$

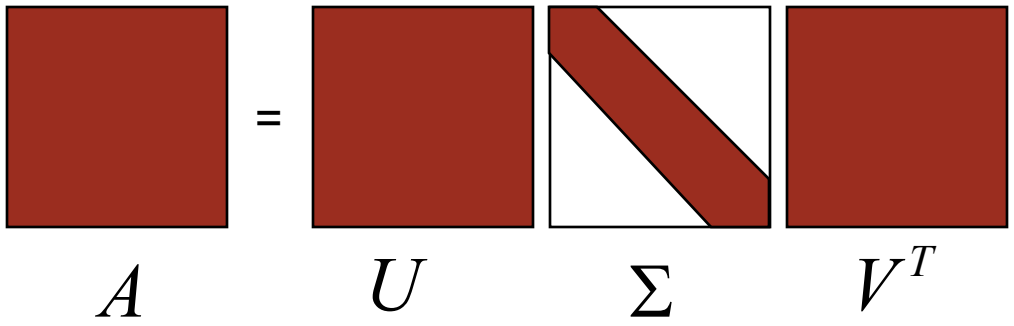
- $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$: Eigenvalues of A
- $u_1, u_2, \ldots, u_r$: Eigenvectors of A

SINGULAR VALUE DECOMPOSITION

- The left singular vectors are an orthonormal basis for the row space of **A**.
- The right singular vectors are an orthonormal basis for the column space of **A**.
- If **A** has rank **r**, then **A** can be written as the sum of **r** rank-1 matrices
- There are **r** "linear components" (trends) in **A**.
  - Linear trend: the tendency of the row vectors of **A** to align with vector **v**
  - Strength of the i-th linear trend: $||Av_i|| = \sigma_i$

SVD MORE FORMALLY

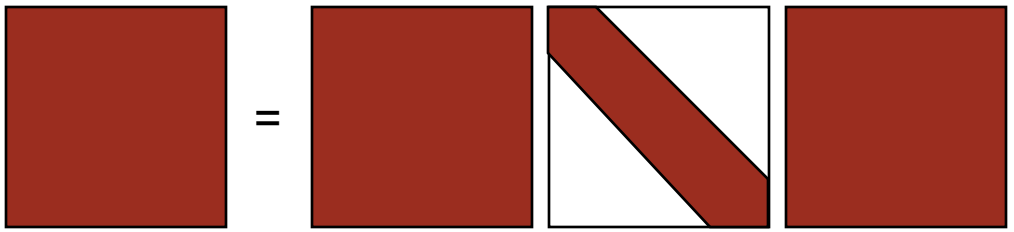
- SVD exists for any matrix
- Formal definition:
  - For square matrices $A \in R^{n \times n}$, there exist orthogonal matrices $U, V \in R^{n \times n}$ and a diagonal matrix Σ, such that all the diagonal values σ_i of Σ are non-negative and



$$A = U \Sigma V^T$$

SVD MORE FORMALLY

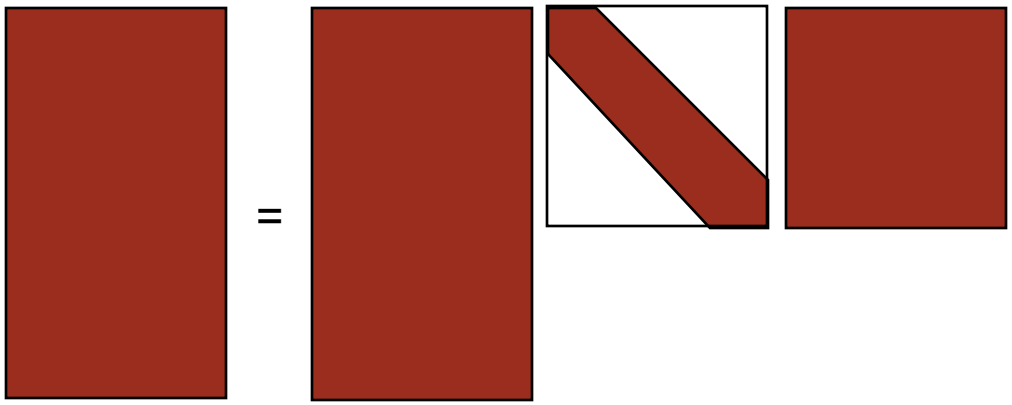
- The diagonal values of Σ ($\sigma_1, \ldots, \sigma_n$) are called the singular values. It is accustomed to sort them: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$
- The columns of U ($\mathbf{u}_1, \ldots, \mathbf{u}_n$) are called the left singular vectors. They are the axes of the ellipsoid.
- The columns of V ($\mathbf{v}_1, \ldots, \mathbf{v}_n$) are called the right singular vectors. They are the preimages of the axes of the ellipsoid.



$$A = U\Sigma V^T$$

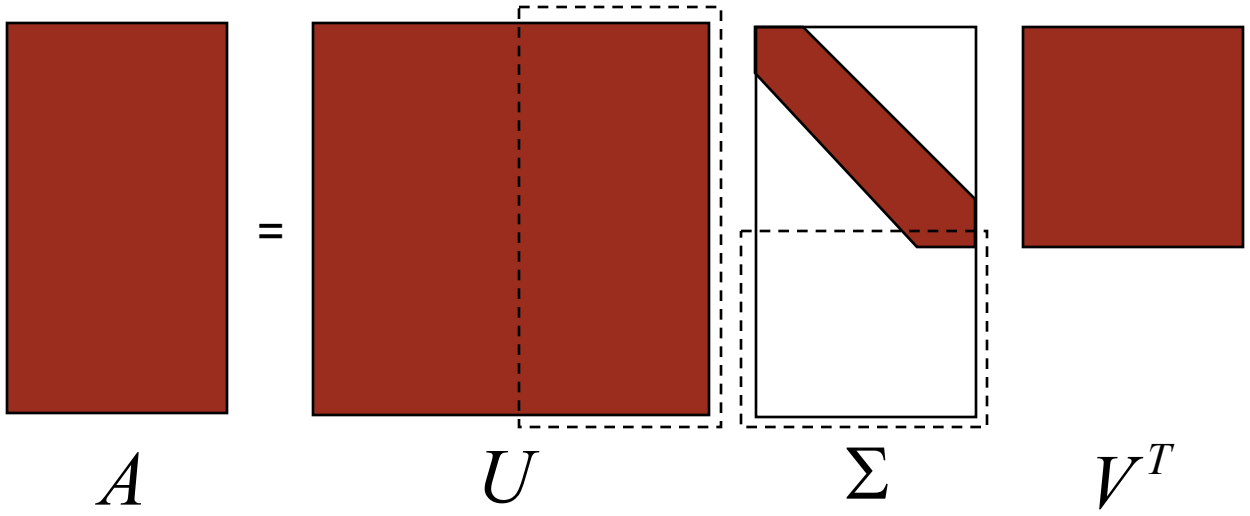
REDUCED SVD

- For rectangular matrices, we have two forms of SVD. The reduced SVD looks like this:
  - The columns of U are orthonormal
  - Cheaper form for computation and storage



FULL SVD

- We can complete U to a full orthogonal matrix and pad Σ by zeros accordingly

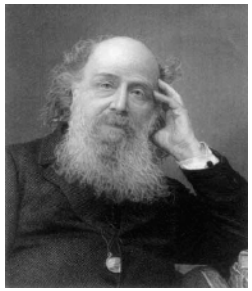


SOME HISTORY

- SVD was discovered by the following people:

E. Beltrami
(1835 – 1900)

M. Jordan
(1838 – 1922)



J. Sylvester
(1814 – 1897)

E. Schmidt
(1876-1959)

H. Weyl
(1885-1955)

SVD IS THE "WORKING HORSE" OF LINEAR ALGEBRA

- There are numerical algorithms to compute SVD. Once you have it, you have many things:
  - Matrix inverse $\rightarrow$ can solve square linear systems
  - Numerical rank of a matrix
  - Can solve least-squares systems
  - PCA
  - Many more…

MATRIX INVERSE AND SOLVING LINEAR SYSTEMS

$$A = U \Sigma V^T$$

- Matrix inverse:

$$A^{-1} = \left(U \Sigma V^T\right)^{-1} = \left(V^T\right)^{-1} \Sigma^{-1} U^{-1} =$$

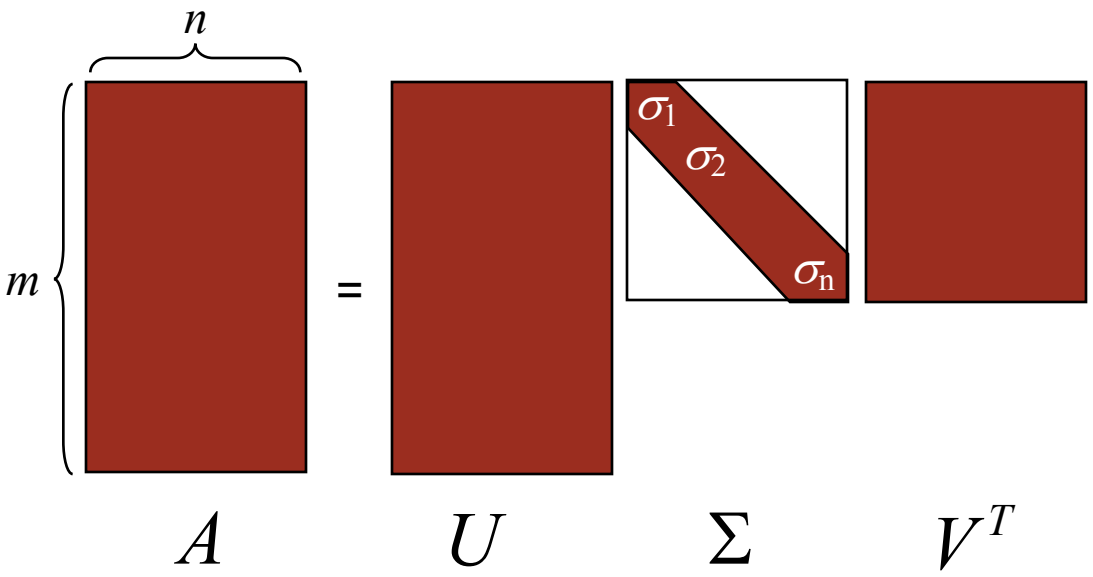
$$= V \begin{bmatrix} \frac{1}{\sigma_1} & & \\ & \ddots & \\ & & \frac{1}{\sigma_n} \end{bmatrix} U^T$$

- So, to solve $A\mathbf{x} = \mathbf{b}$

$$\mathbf{x} = V \Sigma^{-1} U^T \mathbf{b}$$

MATRIX RANK

- The rank of A is the number of non-zero singular values



NUMERICAL RANK

- If there are very small singular values, then A is close to being singular. We can set a threshold t, so that

  $\text{numeric\_rank}(A) = \#\{\sigma_i | \ \sigma_i > t\}$

- If $\text{rank}(A) < n$ then A is singular. It maps the entire space R^n onto some subspace, like a plane (so A is some sort of projection).

RANK-K APPROXIMATIONS (A_K)

$$\left(A_k \right) = \left(U_k \right) \cdot \left(\Sigma_k \right) \cdot \left(V_k^T \right)$$

n x d n x k k x k k x d

U_k (V_k): orthogonal matrix containing the top ***k*** left (right) singular vectors of **A**.

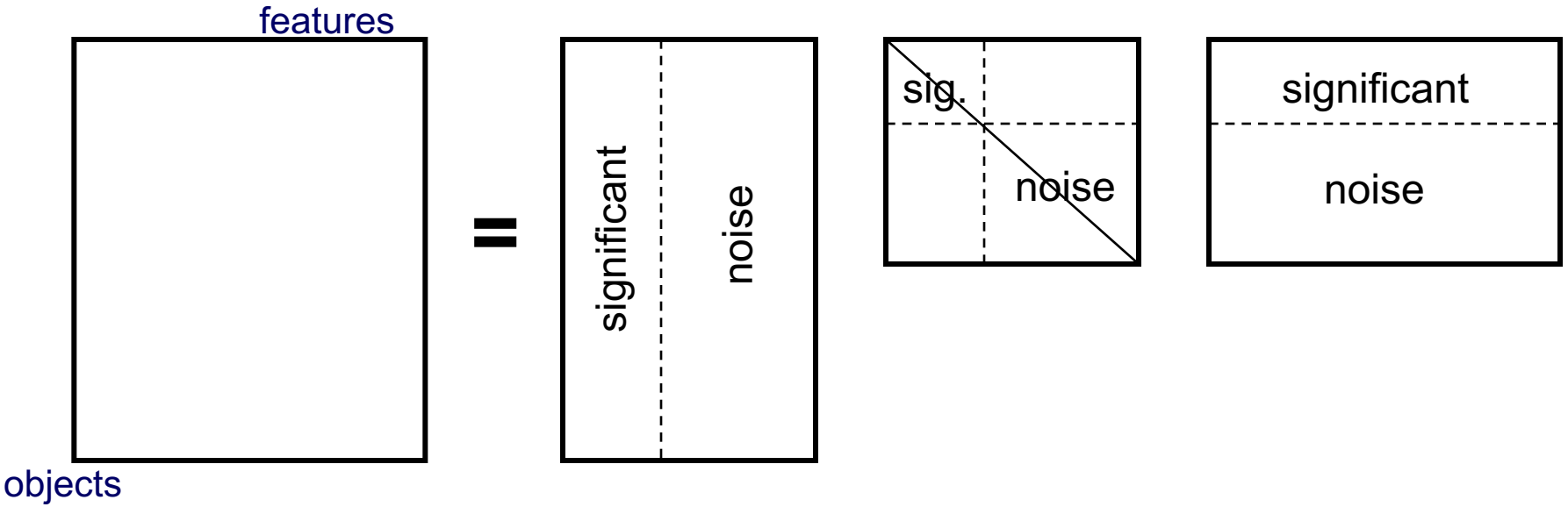
Σ_k: diagonal matrix containing the top ***k*** singular values of **A**

A_k is an approximation of **A**

A_k is the **best** approximation of **A**

SVD AND RANK-K APPROXIMATIONS

$$A = U \Sigma V^T$$



SVD AS AN OPTIMIZATION

- The rank-k approximation matrix A_k produced by the top-k singular vectors of A minimizes the Frobenious norm of the difference with the matrix A

$$A_k = \arg \max_{B:rank(B)=k} \|A - B\|_F^2$$

$$\|A - B\|_F^2 = \sum_{i,j} \left(A_{ij} - B_{ij}\right)^2$$

WHAT DOES THIS MEAN?

- We can project the row (and column) vectors of the matrix A into a k-dimensional space and preserve most of the information
- (Ideally) The k dimensions reveal latent features/aspects/topics of the term (document) space.
- (Ideally) The A_k approximation of matrix A, contains all the useful information, and what is discarded is noise

LATENT FACTOR MODEL

- Rows (columns) are linear combinations of k latent factors
  - E.g., in our extreme document example there are two factors
- Some noise is added to this rank-k matrix resulting in higher rank

- SVD retrieves the latent factors (hopefully).

SVD AND PCA

- PCA is a special case of SVD on the centered covariance matrix.

COVARIANCE MATRIX

- Goal: reduce the dimensionality while preserving the "information in the data"
- Information in the data: variability in the data
  - We measure variability using the covariance matrix.
  - Sample covariance of variables X and Y

$$\sum_i (x_i - \mu_X)^T (y_i - \mu_Y)$$

- Given matrix A, remove the mean of each column from the column vectors to get the centered matrix C
- The matrix $V = C^T C$ is the covariance matrix of the row vectors of A.

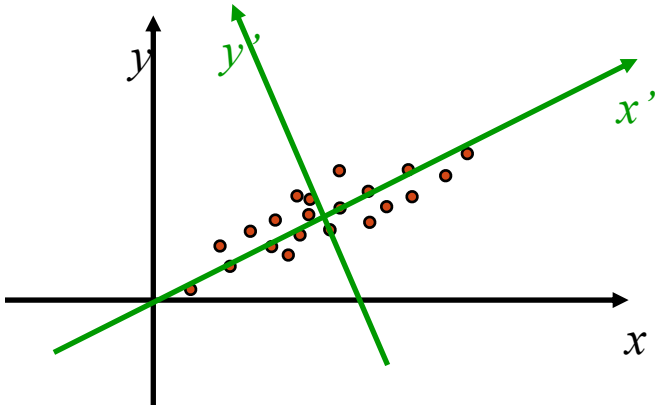
PCA: PRINCIPAL COMPONENT ANALYSIS

- We will project the rows of matrix A into a new set of attributes (dimensions) such that:
  - The attributes have zero covariance to each other (they are orthogonal)
  - Each attribute captures the most remaining variance in the data, while orthogonal to the existing attributes
    - The first attribute should capture the most variance in the data

- For matrix C, the variance of the rows of C when projected to vector x is given by $\sigma^2 = \left||Cx|\right|^2$
  - The right singular vector of C maximizes σ^2!

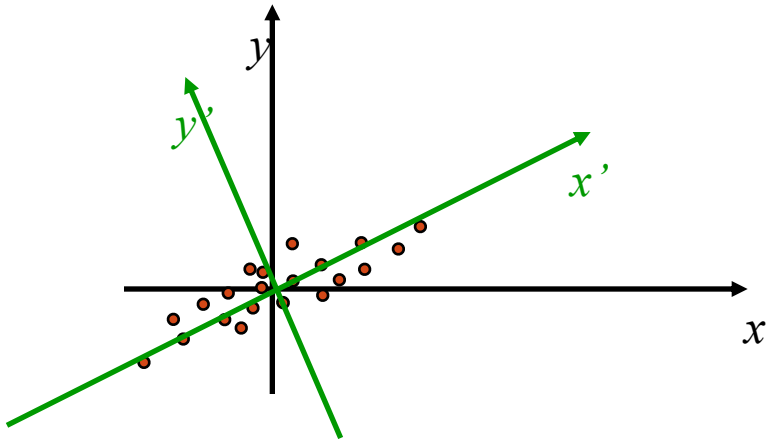
PCA

- We wanted to find principal components



PCA

- Move the center of mass to the origin
  - $\mathbf{p}_i' = \mathbf{p}_i - \mathbf{m}$



PCA

- Constructed the matrix X of the data points.

$$X = \begin{bmatrix} | & | & & | \\ \mathbf{p}'_1 & \mathbf{p}'_2 & \cdots & \mathbf{p}'_n \\ | & | & & | \end{bmatrix}$$

- The principal axes are eigenvectors of $S = XX^T$

$$XX^T = S = U \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} U^T$$

PCA

- We can compute the principal components by SVD of X:

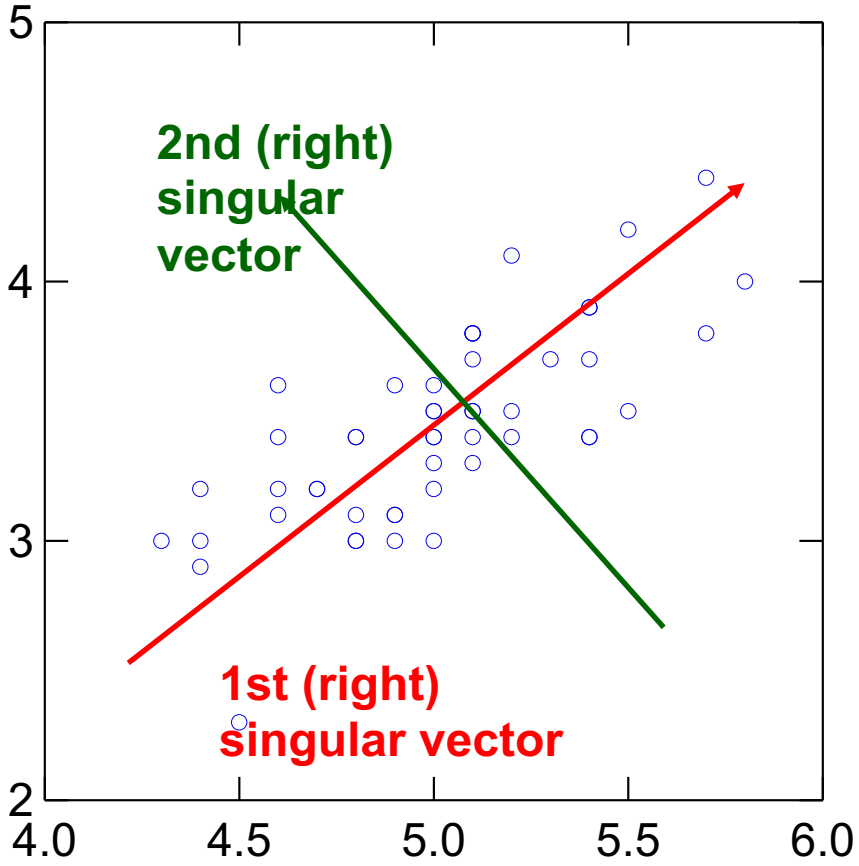
$$X = U\Sigma V^T$$

$$XX^T = U\Sigma V^T (U\Sigma V^T)^T =$$

$$= U\Sigma V^T V\Sigma^T U^T = U\tilde{\Sigma}^2 U^T$$

- Thus, the left singular vectors of X are the principal components! We sort them by the size of the singular values of X.

PCA



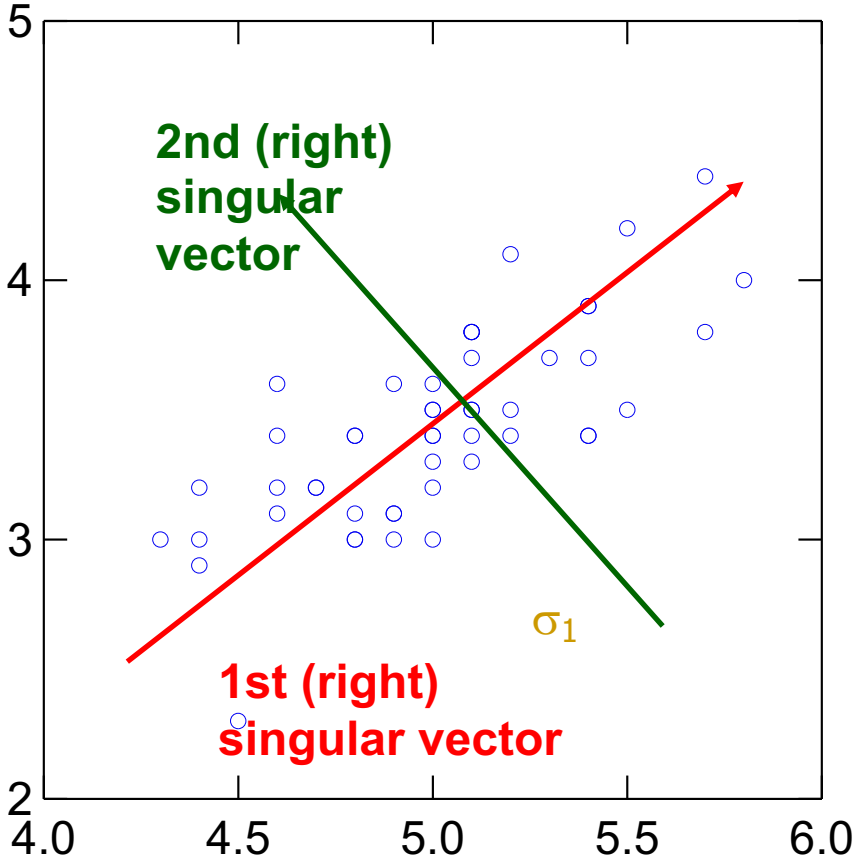
Input: 2-d dimensional points

Output:

1st (right) singular vector: direction of maximal variance,

2nd (right) singular vector: direction of maximal variance, after removing the projection of the data along the first singular vector.

SINGULAR VALUES



σ_1: measures how much of the data variance is explained by the first singular vector.

σ_2: measures how much of the data variance is explained by the second singular vector.

SINGULAR VALUES TELL US SOMETHING ABOUT THE VARIANCE

- The variance in the direction of the **k**-th principal component is given by the corresponding singular value σ_k^2

- Singular values can be used to estimate how many components to keep

- ***Rule of thumb:*** keep enough to explain ***85%*** of the variation:

$$\frac{\sum_{j=1}^{k} \sigma_j^2}{\sum_{j=1}^{n} \sigma_j^2} \approx 0.85$$

EXAMPLE

drugs

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

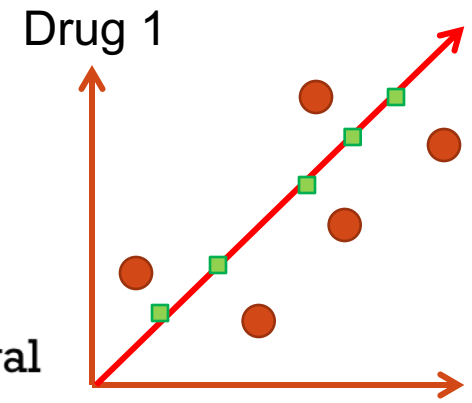
students

legal illegal

a_{ij}: usage of student i of drug j

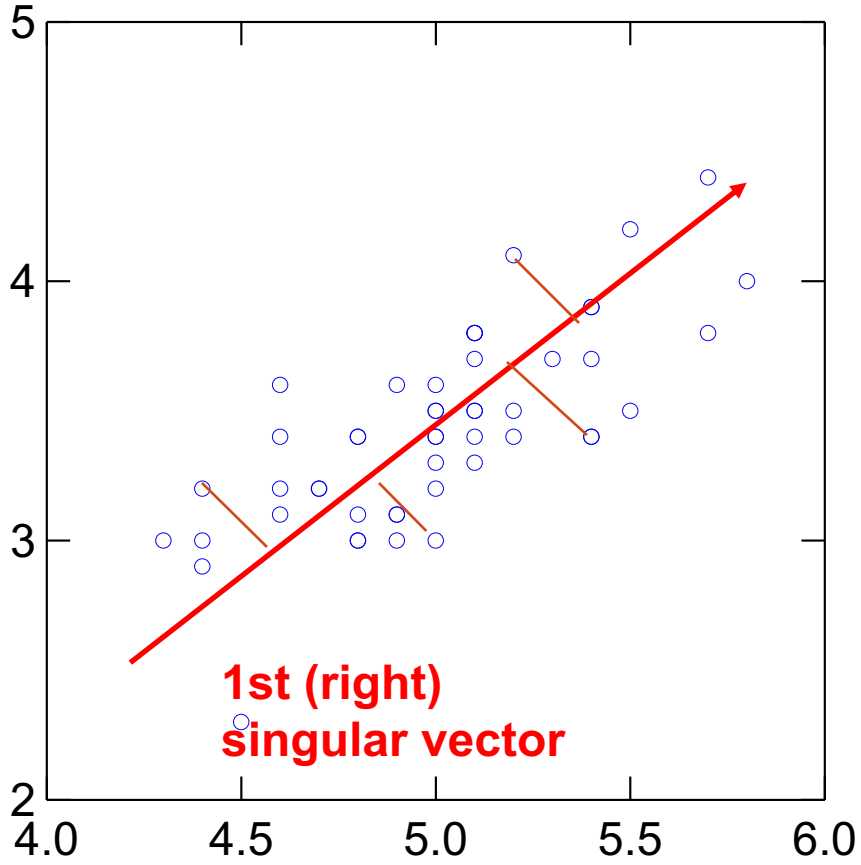
$$A = U\Sigma V^T$$

- First right singular vector v_1
  - More or less same weight to all drugs
  - Discriminates heavy from light users
- Second right singular vector
  - Positive values for legal drugs, negative for illegal



ANOTHER PROPERTY OF PCA/SVD

- The chosen vectors are such that minimize the sum of square differences between the data vectors and the low-dimensional projections



SVD is “**the Rolls-Royce and the Swiss Army Knife of Numerical Linear Algebra.**”*

*Dianne O’Leary, MMDS ’06