SVD Application: Categorizing Movie Data

Laura Lyman Margot Gerritsen

Stanford University

ICME Summer Workshops
Fundamentals of Data Science

August 2, 2021

Simple example with movie data

Let A be a rectangular data matrix, whose

- rows correspond to users,
- columns correspond to movies, and
- entries are ratings that users have given movies.

Specifically, A_{ij} = rating user i has given movie j. For example,

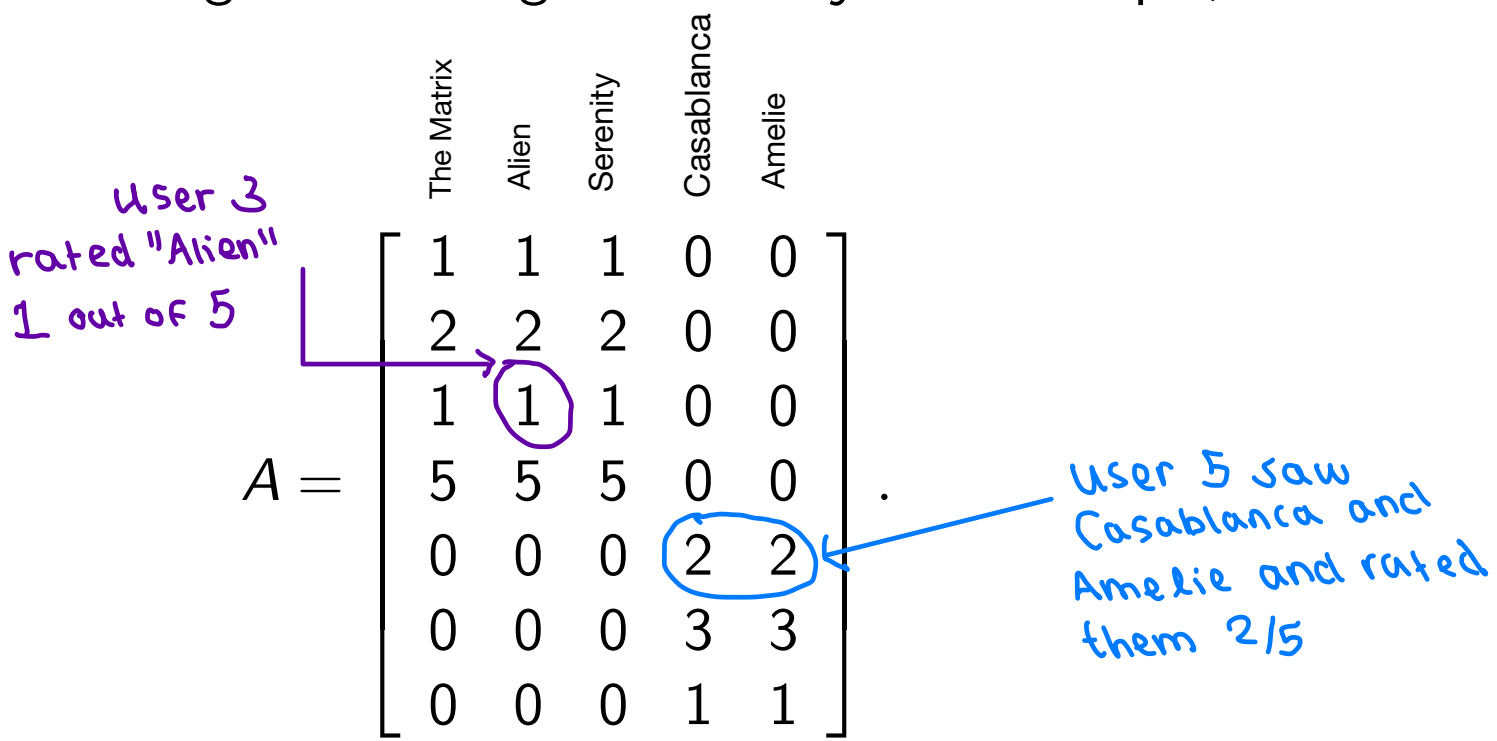
$$A = \begin{array}{c} \begin{array}{ccccc} \text{The Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array}.$$

Simple example with movie data

Let A be a rectangular data matrix, whose

- rows correspond to users,
- columns correspond to movies, and
- entries are ratings that users have given movies.

Specifically, A_{ij} = rating user i has given movie j. For example,



Simple example with movie data

Let A be a rectangular data matrix, whose

- rows correspond to users,
- columns correspond to movies, and
- entries are ratings that users have given movies.

Specifically, A_{ij} = rating user i has given movie j. For example,

$A_{ij} = 0 \iff$ user i has not rated movie j

$$A = \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \text{The Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

e.g. first 4 users didn't rate Casablanca or Amelie

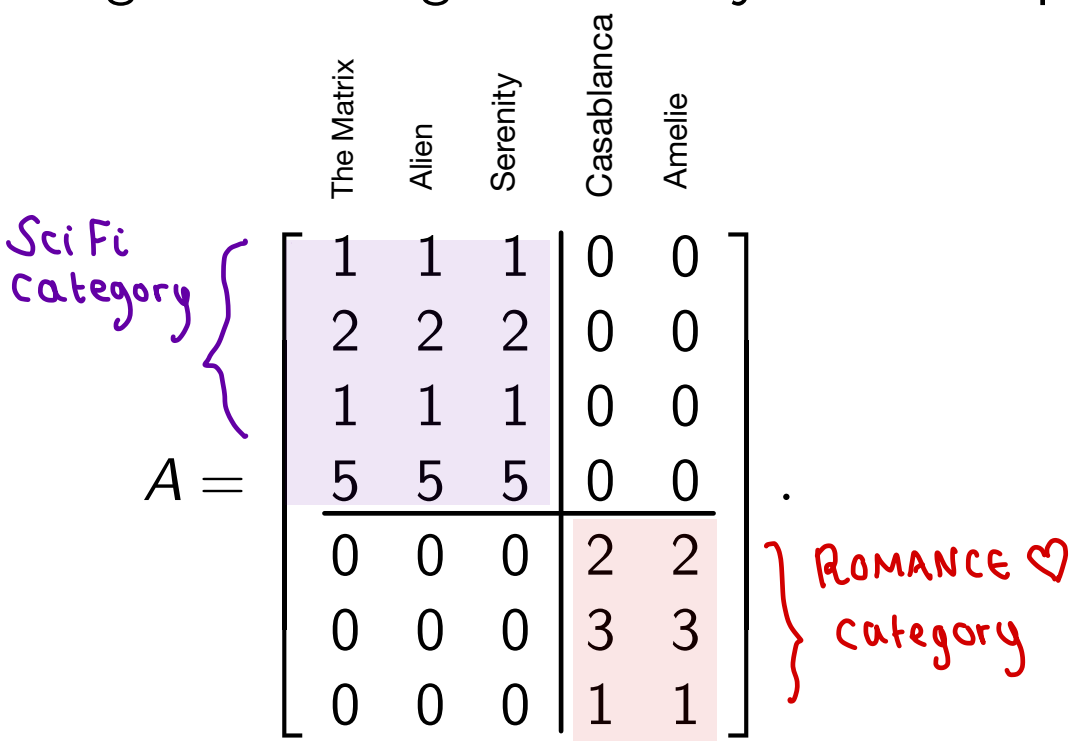
last 3 users only

Simple example with movie data

Let A be a rectangular data matrix, whose

- rows correspond to users,
- columns correspond to movies, and
- entries are ratings that users have given movies.

Specifically, A_{ij} = rating user i has given movie j. For example,



SVD for movie data

Using the compact SVD,

$$\underbrace{\left[\begin{array}{ccc|cc} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{array}\right]}_{\text{SCI FI / Romance}} = \overbrace{\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}}^{U} \times \overbrace{\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}}^{\Sigma} \times \overbrace{\begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}^{T}}^{V^T}.$$

- Σ: # of categories and their relative strength in the data set

SVD for movie data

Using the compact SVD,

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

SCI FI

Romance

if user 3 rates higher

$$U \begin{bmatrix} 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

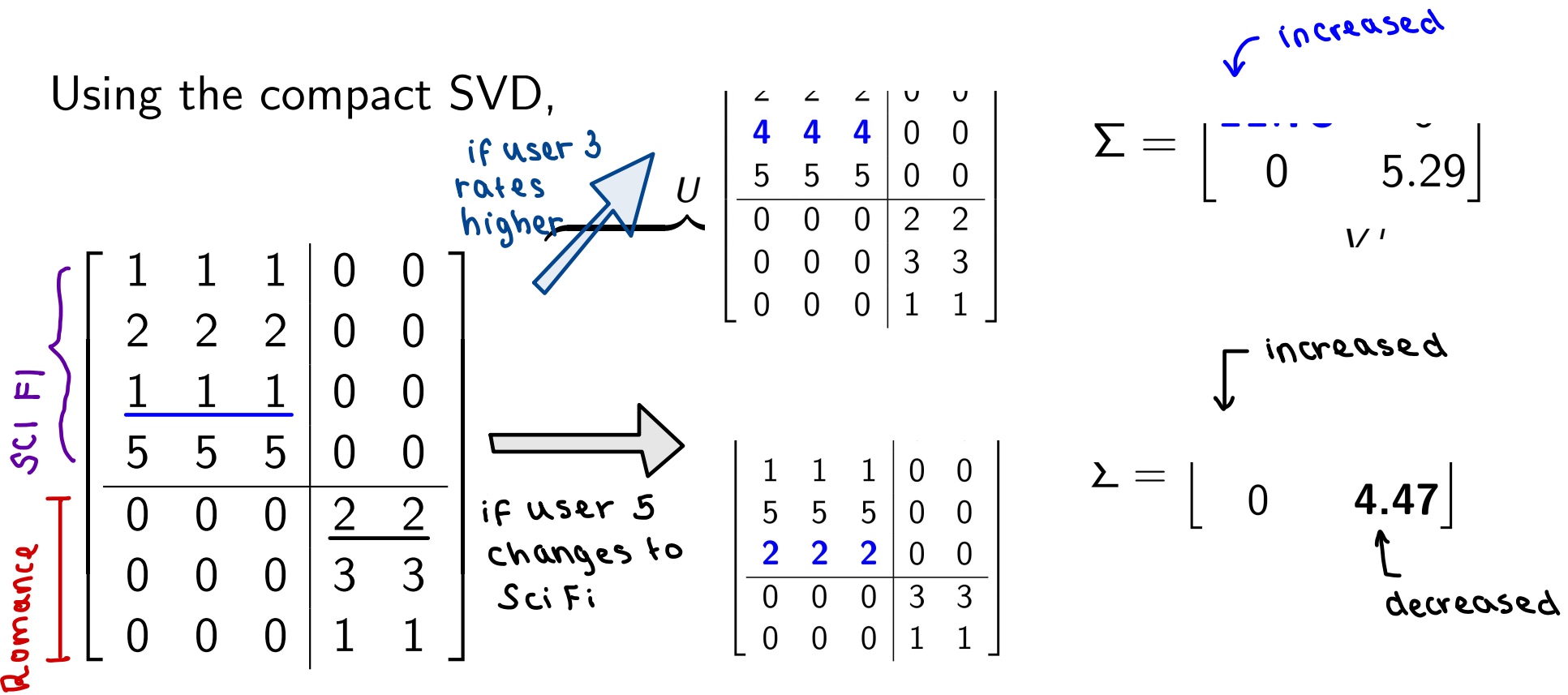
$$\Sigma = \begin{bmatrix} & \\ 0 & 5.29 \end{bmatrix}$$

V'

- Σ: # of categories and their relative strength in the data set

SVD for movie data

Using the compact SVD,



- Σ: # of categories and their relative strength in the data set

SVD for movie data

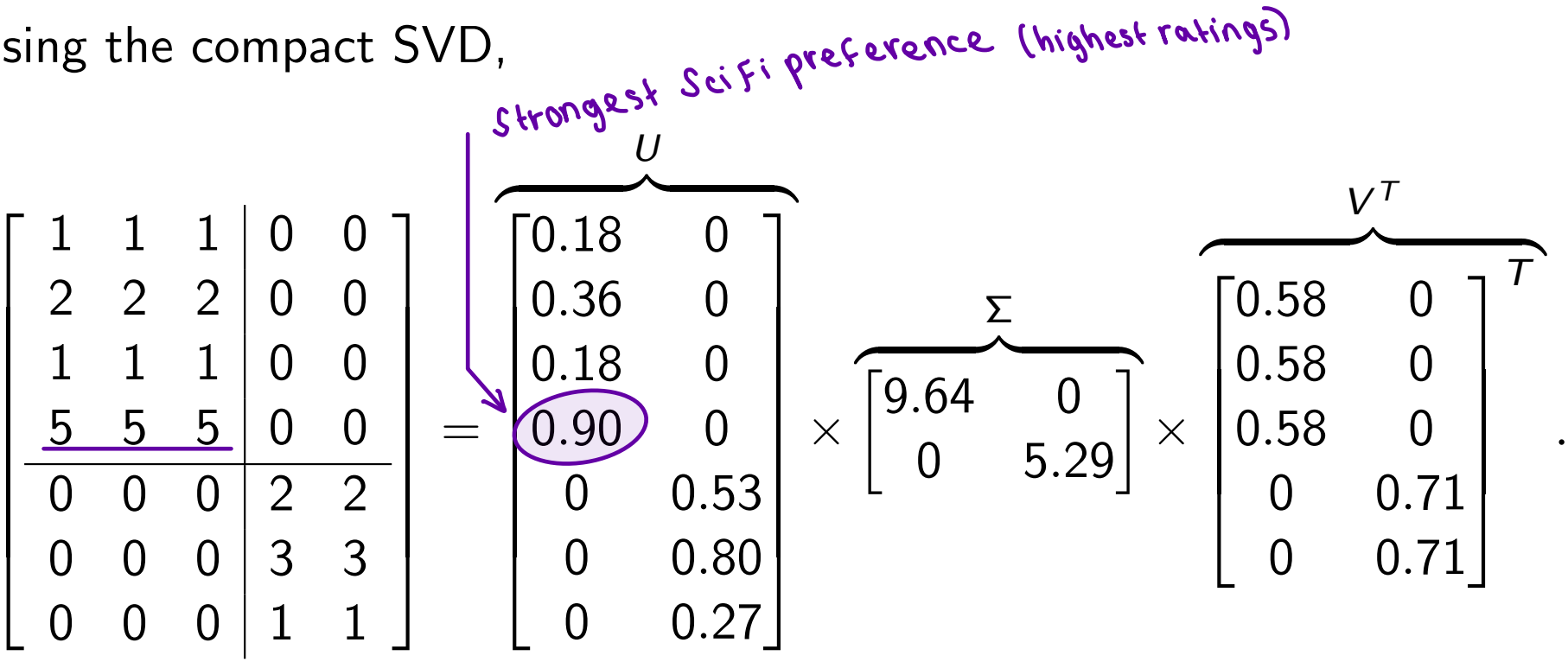
Using the compact SVD,

$$\left[\begin{array}{ccc|cc} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{array}\right] = \overbrace{\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}}^{U} \times \overbrace{\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}}^{\Sigma} \times \overbrace{\begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}^{T}}^{V^T} .$$

- Σ: # of categories and their relative strength in the data set
- U: "user to category (genre)" similarity matrix

SVD for movie data

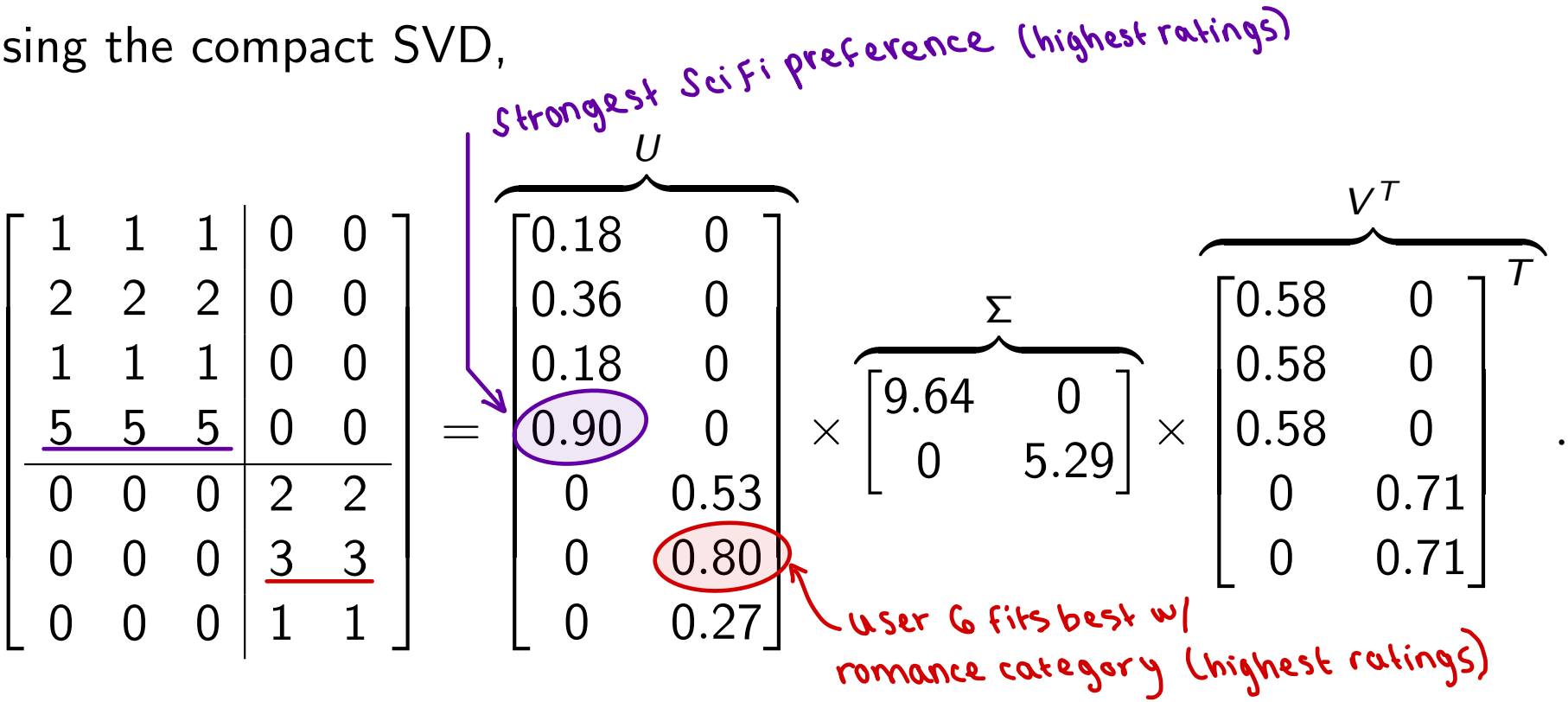
Using the compact SVD,



- Σ: # of categories and their relative strength in the data set
- U: "user to category (genre)" similarity matrix

SVD for movie data

Using the compact SVD,



$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}^T .$$

- Σ: # of categories and their relative strength in the data set
- U: "user to category (genre)" similarity matrix

SVD for movie data

Using the compact SVD,

each user gives identical ratings

first 3 movies belong to Sci Fi to the same extent (and are separate from romance)

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \overbrace{\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}}^{U} \times \overbrace{\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}}^{\Sigma} \times \overbrace{\begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}^{T}}^{V^T} .$$

- Σ: # of categories and their relative strength in the data set
- U: "user to category (genre)" similarity matrix
- V: "category (genre) to movie" similarity matrix

SVD for movie data

Using the compact SVD,

each user gives identical ratings

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \overbrace{\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}}^{U} \times \overbrace{\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}}^{\Sigma} \times \overbrace{\begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}^{T}}^{V^T}$$

all rate Casablanca and Amelie equally

first 3 movies belong to Sci Fi to the same extent (and are separate from romance)

belong equally to romance ♡

- Σ: # of categories and their relative strength in the data set
- U: "user to category (genre)" similarity matrix
- V: "category (genre) to movie" similarity matrix

Diversifying within a category

Suppose instead the second user — let's call him Pete — has non-homogenous ratings; he really prefers Alien to the other Sci Fi movies and rates it a 5.

Diversifying within a category

Suppose instead the second user — let's call him Pete — has non-homogenous ratings; he really prefers Alien to the other Sci Fi movies and rates it a 5.

- Now 3 nonzero singular values
- In compact SVD, we now need an extra column in U and V^T to span the Sci-fi columns of A (which are no longer identical)

$$\left[\begin{array}{ccc|cc} 1 & 1 & 1 & 0 & 0 \\ 2 & \mathbf{5} & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{array}\right] = \overbrace{\begin{bmatrix} \mathbf{0.16} & 0 & \mathbf{0.10} \\ \mathbf{0.52} & 0 & \mathbf{-0.85} \\ \mathbf{0.16} & 0 & \mathbf{0.10} \\ \mathbf{0.82} & 0 & \mathbf{0.50} \\ 0 & 0.53 & 0 \\ 0 & 0.80 & 0 \\ 0 & 0.27 & 0 \end{bmatrix}}^{U} \overbrace{\begin{bmatrix} \mathbf{10.47} & 0 & 0 \\ 0 & 5.29 & 0 \\ 0 & 0 & \mathbf{2.11} \end{bmatrix}}^{} \begin{bmatrix} \mathbf{0.52} & \mathbf{0.67} & \mathbf{0.52} & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \\ \mathbf{0.48} & \mathbf{-0.74} & \mathbf{0.48} & 0 & 0 \end{bmatrix}$$

Diversifying within a category

Suppose instead the second user — let's call him Pete — has non-homogenous ratings; he really prefers Alien to the other Sci Fi movies and rates it a 5.

- Now 3 nonzero singular values
- In compact SVD, we now need an extra column in U and V^T to span the Sci-fi columns of A (which are no longer identical)

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & \mathbf{5} & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \overbrace{\begin{bmatrix} \mathbf{0.16} & 0 & \mathbf{0.10} \\ \mathbf{0.52} & 0 & \mathbf{-0.85} \\ \mathbf{0.16} & 0 & \mathbf{0.10} \\ \mathbf{0.82} & 0 & \mathbf{0.50} \\ 0 & 0.53 & 0 \\ 0 & 0.80 & 0 \\ 0 & 0.27 & 0 \end{bmatrix}}^{U} \overbrace{\begin{bmatrix} \mathbf{10.47} & 0 & 0 \\ 0 & 5.29 & 0 \\ 0 & 0 & \mathbf{2.11} \end{bmatrix}}$$

$$\begin{bmatrix} \mathbf{0.52} & \mathbf{0.67} & \mathbf{0.52} & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \\ \mathbf{0.48} & \mathbf{-0.74} & \mathbf{0.48} & 0 & 0 \end{bmatrix}$$

≈ unchanged values (romance category)

Mixing categories

Instead of updating ratings within a category, what if our second user (Pete) is interested in Sci-fi *and* romance? In fact, he *loves* Amelie.

Mixing categories

Instead of updating ratings within a category, what if our second user (Pete) is interested in Sci-fi *and* romance? In fact, he *loves* Amelie.

- Now there is "fill-in" where there were previously 0s when diversified ratings still fell into separate categories

Mixing categories

Instead of updating ratings within a category, what if our second user (Pete) is interested in Sci-fi *and* romance? In fact, he *loves* Amelie.

- Now there is "fill-in" where there were previously 0s when diversified ratings still fell into separate categories
- The magnitude of the "fill-in" corresponds to the strength of the new connection i.e. the mixed Sci-fi-romance grouping

Mixing categories

Instead of updating ratings within a category, what if our second user (Pete) is interested in Sci-fi *and* romance? In fact, he *loves* Amelie.

- Now there is "fill-in" where there were previously 0s when diversified ratings still fell into separate categories
- The magnitude of the "fill-in" corresponds to the strength of the new connection i.e. the mixed Sci-fi-romance grouping
- A weaker connection (say $A_{2,5} = 2$ instead of 4) gives a smaller corresponding Σ_{33}, along with fill-in entries smaller in magnitude

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & \mathbf{4} \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.17 & 0.06 & -0.06 \\ 0.43 & -0.44 & 0.79 \\ 0.17 & 0.06 & -0.06 \\ 0.86 & 0.29 & -0.31 \\ 0.05 & -0.45 & -0.28 \\ 0.08 & -0.68 & -0.42 \\ 0.03 & -0.23 & -0.14 \end{bmatrix} \begin{bmatrix} 9.80 & 0 & 0 \\ 0 & 5.97 & 0 \\ 0 & 0 & 2.30 \end{bmatrix} \begin{bmatrix} 0.56 & 0.11 & -0.05 \\ 0.56 & 0.11 & -0.05 \\ 0.56 & 0.11 & -0.05 \\ 0.04 & -0.53 & -0.85 \\ 0.21 & -0.82 & 0.53 \end{bmatrix}^T$$

= fill in from mixing categories