Machine Learning & Data Mining
Miniproject 2: MovieLens Visualization

Caltech CS/CNS/EE 155
Released February 16^{st}, 2022

---

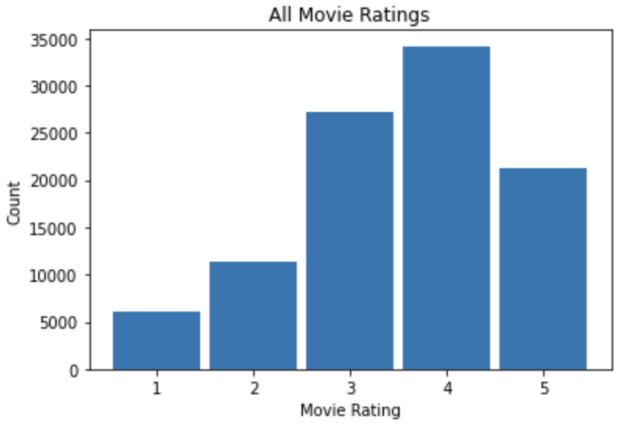
1 Introduction

- Team name: Darth Jar Jar
- Members: Ben Juarez, Kyle McGraw, Dallas Taylor
- Piazza Post : https://piazza.com/class/kxhtyed0nmh2s?cid=369
- Work Division:
  - Each team member was involved in each section, and the work load was distributed equally. Dallas did the majority of the basic visualizations. Ben did the majority of the implementations of methods 1 and 2 for finding U and V. Kyle did the majority of the implementation of method 3 for finding U and V. The whole team collaborated to generate the visualizations for each method. Furthermore, the report workload was distributed equally.
- Packages:
  - numpy
  - matplotlib - pyplot
  - pandas
  - csv
  - random
  - surpise - SVD, Dataset, Reader, accuracy
  - surpise.model_selection - cross_validate

2 Basic Visualizations

Question 2 Code

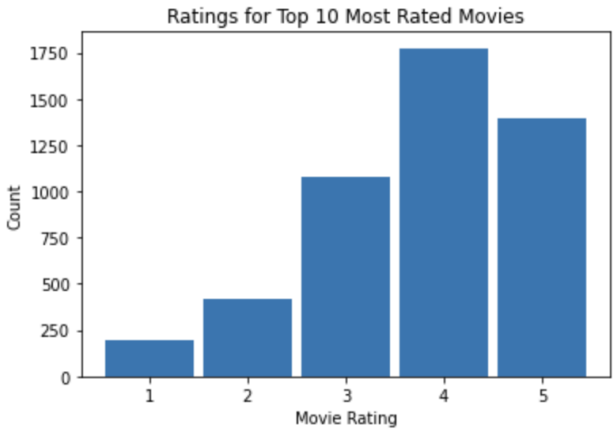
1. All ratings in the MovieLens Dataset.



General Observations The data is skewed right such that the mode rating is 4 and the ratings of 1 and 2 have the lowest counts of all ratings.

Results vs Expectation The results are similar to as expected, where we would assume most ratings to be mostly positive, with lower counts of the highest and lowest ratings 1 and 5.

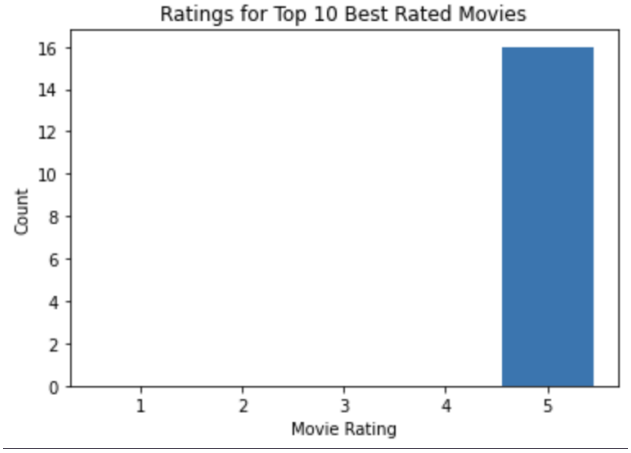
2. All ratings of the ten most popular movies (movies which have received the most ratings).



General Observations The distribution of this histogram is very similar to that of the entire MovieLens Dataset. There are higher counts of reviews of 4 and 5 and lower counts of reviews of 1 and 2. Additionally, we see that approximately 5000 of the reviews account for the 10 most reviewed movies.

Results vs Expectation The distribution being similar to all of the data is as expected, since these 10 movies will have the largest impact on the rating distribution, and thus should closely reflect the overall distribution. Additionally, it is as expected that we would have slightly higher counts of 4 and 5 and lower counts of 1 and 2 since the most popular movies would logically also probably have more positive reviews due to mass appeal.

3. All ratings of the ten best movies (movies with the highest average ratings).



General Observations We see that this distribution is simply only a bar with all ratings of value 5.

Results vs Expectation These results are as expected, since the best reviewed movies would most likely have only reviews of 5, especially since we are not considering a lower limit on number of reviews for each movie.

Compare to Most Popular Movies This distribution is very different from that of the most popular movies, since the most popular movies utilize a much larger amount of data points (reviews) and thus have larger range in rating value. However, the best reviewed movies only consider movies that have the highest rate, and some movies can be rated only once or twice with a value of 5 to produce a "best rated movie". Thus, for most rated vs best rated, we produce a distribution that mimics the overall distribution vs only a bar at 5.

4. All ratings of movies from three genres of your choice (create three separate visualizations).

General Observations All three distributions closely match the overall distribution. However, we see that the gap between the number of ratings of 3 and 4 relative to each other is closest for the Horror genre. Additionally, Sci-Fi has the largest number of reviews by far, especially considering that the number of reviews with a 4 rating (approx. 4000) for Sci-Fi exceeds the total number of reviews for both of the other genres, separately.

Results vs Expectation These distributions are as expected. The review counts agree with the fact that the Sci-Fi genre is a very popular genre with a lot of movies, while the Animated and Horror genres appeal to a slightly smaller audience and maybe involves users less likely to review. Additionally, the rating distributions for Sci-Fi and Horror agree with the idea of low-budget Sci-Fi and Horror movies that are low quality and thus review poorly, and agrees with the idea of an "average" film in such genres that warrants an "average" rating of 3. The Animated genre, however, had higher 5 counts and lower 1 counts, potentially reflecting on how users that watch that genre may be inclined to review more positively (consider children's movies, large fan followings for anime, etc).

3 Matrix Factorization

Question 3 Code

Method 1 - HW5 SVD (No Bias)

Description

Our method 1 was using the same matrix factorization method from Homework 5 in order to find the matrices U and V (such that $Y \simeq U^T$). With $k = 20$, we kept the same parameters and and stopping criteria as used in the homework. With more time, we would have performed random / grid searches to further optimize these aspects. The MovieLens dataset was split into train.csv and test.csv as given, and this method was run on train.csv with its performance assessed on test.csv using the mean regularized-squared error. To note, the same SGD implementation was used from the homework:

$$\Rightarrow \partial_{u_i} = \frac{\lambda}{2}(2u_i) + \frac{1}{2}\sum_j (-2v_j)(y_{ij} - u_i^T v_j)$$

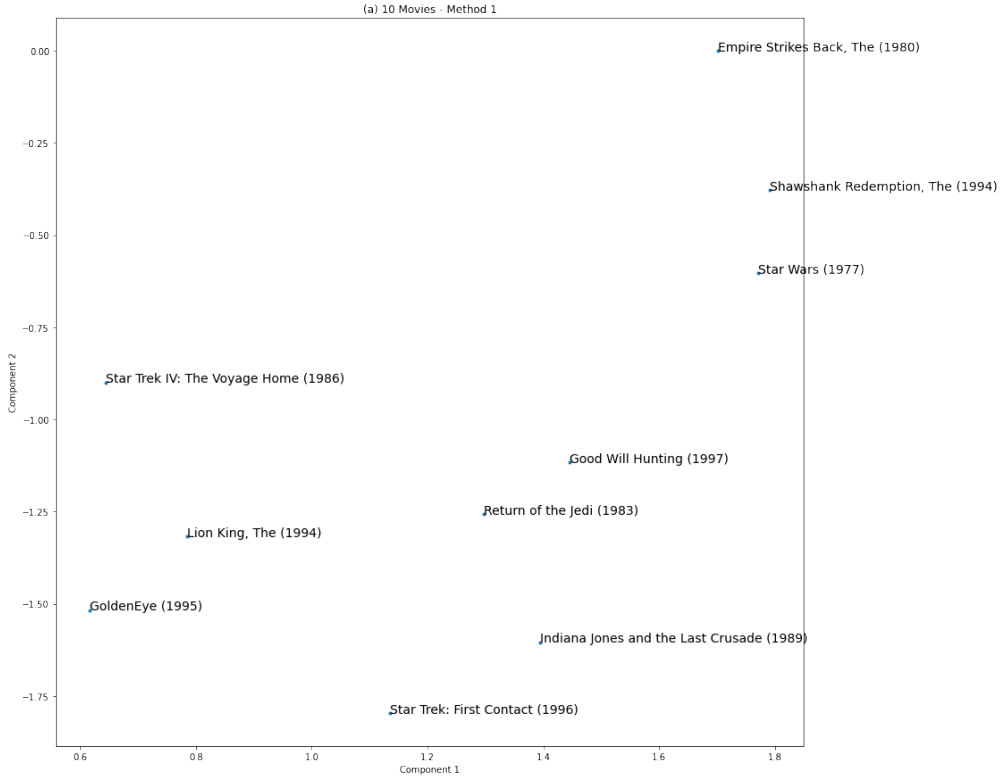
$$\partial_{u_i} = \lambda u_i - \sum_j v_j (y_{ij} - u_i^T v_j)$$

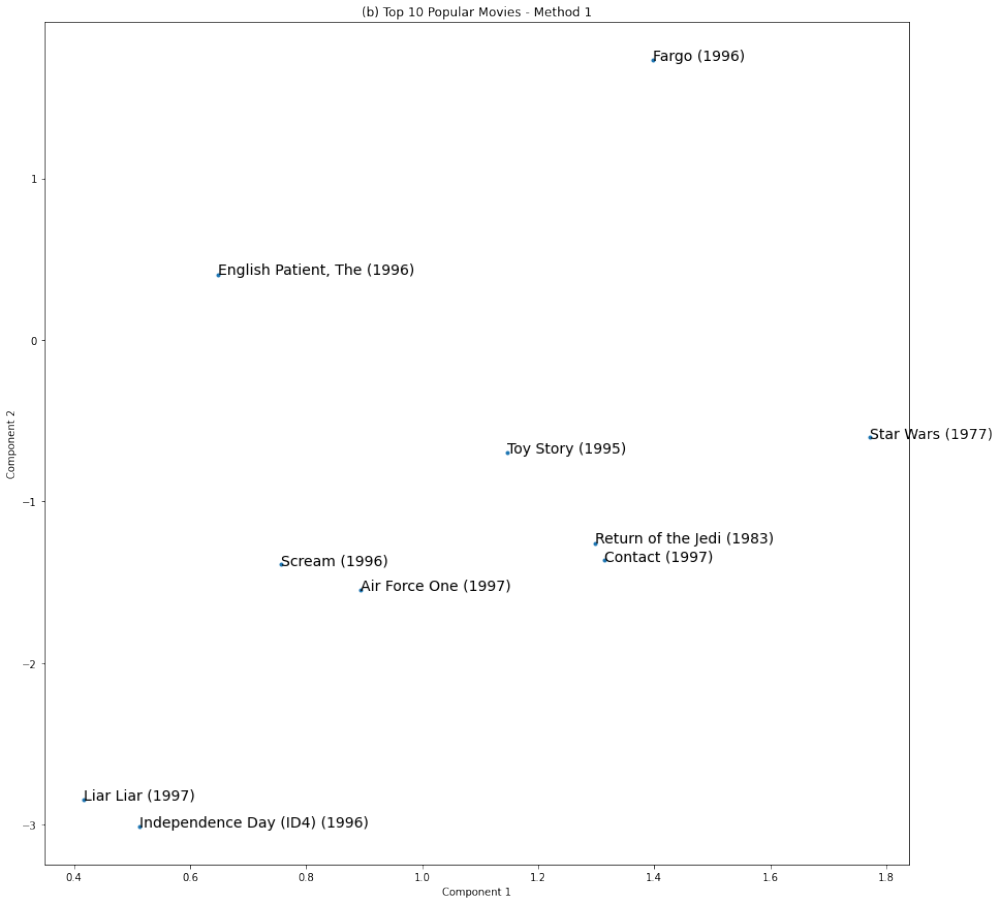
$$\Rightarrow \partial_{v_j} = \lambda v_j - \sum_i u_i (y_{ij} - u_i^T v_j)$$

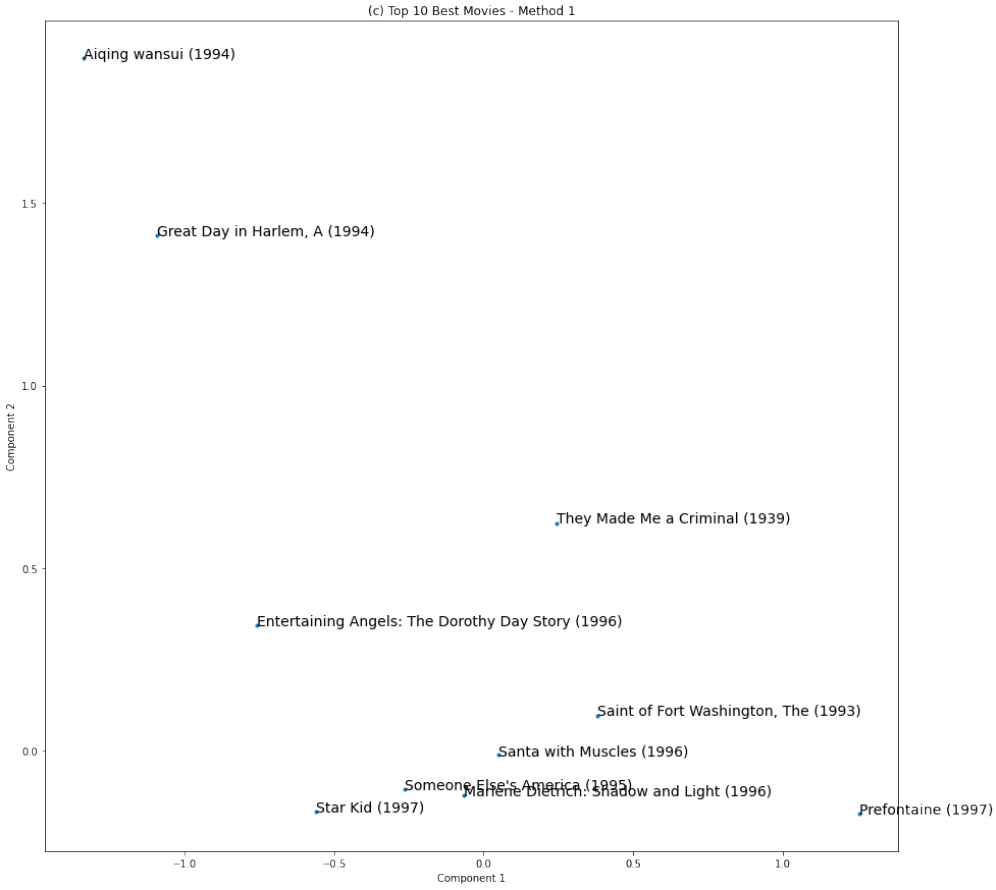
The training objective for this method is as follows:

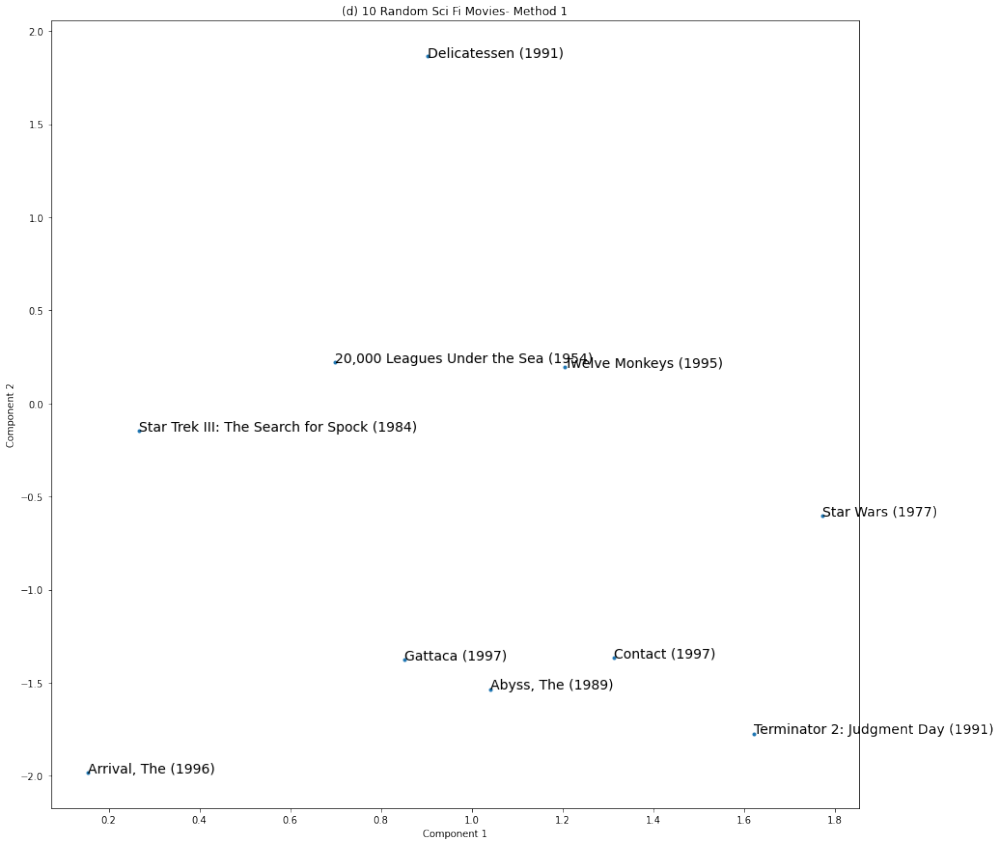
$$\underset{U,V}{\operatorname{argmin}} \frac{\lambda}{2}\left(\|U\|^2 + \|V\|^2\right) + \sum_{(i,j)\in S} \left(Y_{i,j} - u_i^T v_j\right)^2$$

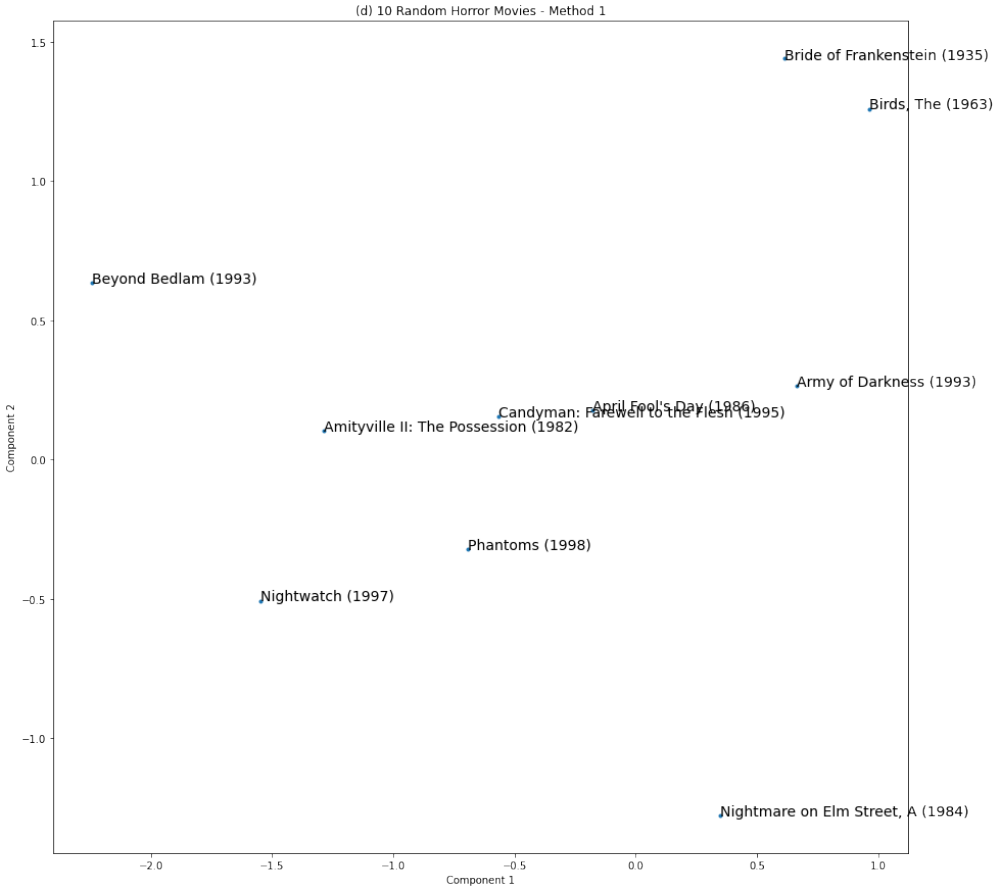
Visualizations

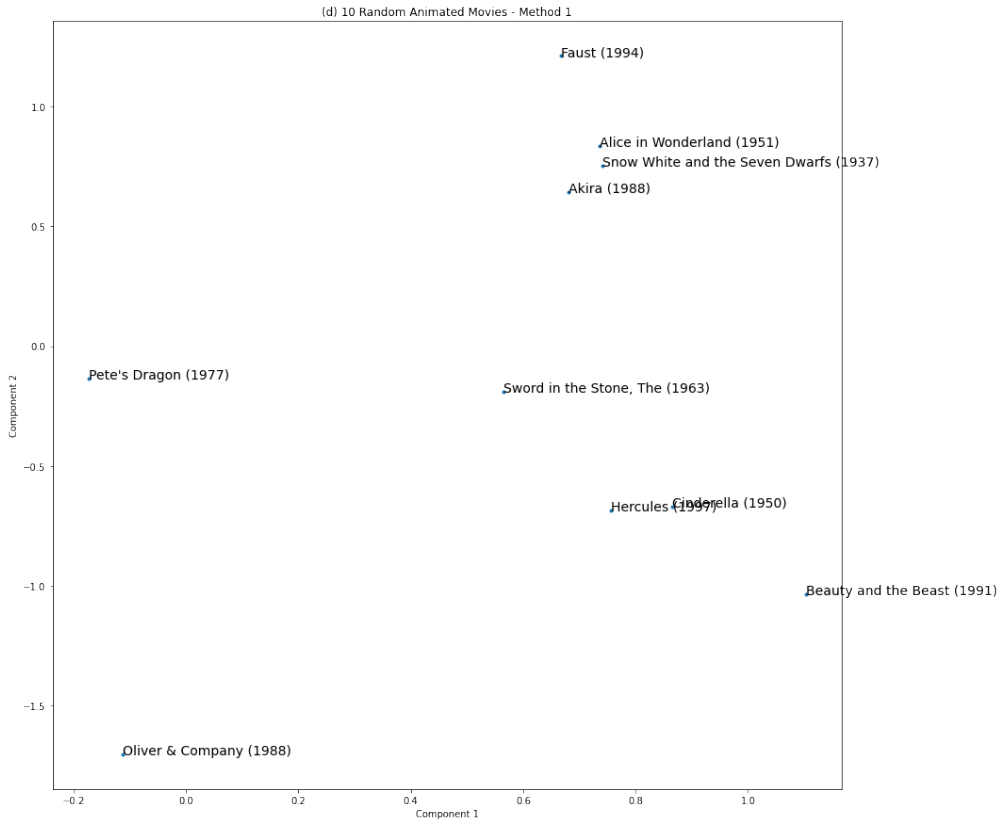



(b) Top 10 Popular Movies - Method 1
Fargo (1996)
English Patient, The (1996)
Toy Story (1995)
Star Wars (1977)
Return of the Jedi (1983)
Contact (1997)
Scream (1996)
Air Force One (1997)
Liar Liar (1997)
Independence Day (ID4) (1996)
Component 1
Component 2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
1
0
−1
−2
−3


(c) Top 10 Best Movies - Method 1
Aiqing wansui (1994)
Great Day in Harlem, A (1994)
They Made Me a Criminal (1939)
Entertaining Angels: The Dorothy Day Story (1996)
Saint of Fort Washington, The (1993)
Santa with Muscles (1996)
Someone Else's America (1995)
Marlene Dietrich: Shadow and Light (1996)
Star Kid (1997)
Prefontaine (1997)
Component 1
Component 2
1.5
1.0
0.5
0.0
−1.0
−0.5
0.0
0.5
1.0


(d) 10 Random Sci Fi Movies- Method 1
Delicatessen (1991)
20,000 Leagues Under the Sea (1954)
Twelve Monkeys (1995)
Star Trek III: The Search for Spock (1984)
Star Wars (1977)
Gattaca (1997)
Contact (1997)
Abyss, The (1989)
Terminator 2: Judgment Day (1991)
Arrival, The (1996)
Component 1
Component 2


(d) 10 Random Horror Movies - Method 1
Bride of Frankenstein (1935)
Birds, The (1963)
Beyond Bedlam (1993)
Army of Darkness (1993)
April Fool's Day (1986)
Candyman: Farewell to the Flesh (1995)
Amityville II: The Possession (1982)
Phantoms (1998)
Nightwatch (1997)
Nightmare on Elm Street, A (1984)
Component 1
Component 2



Method 2 - HW5 SVD (With Bias)

Description

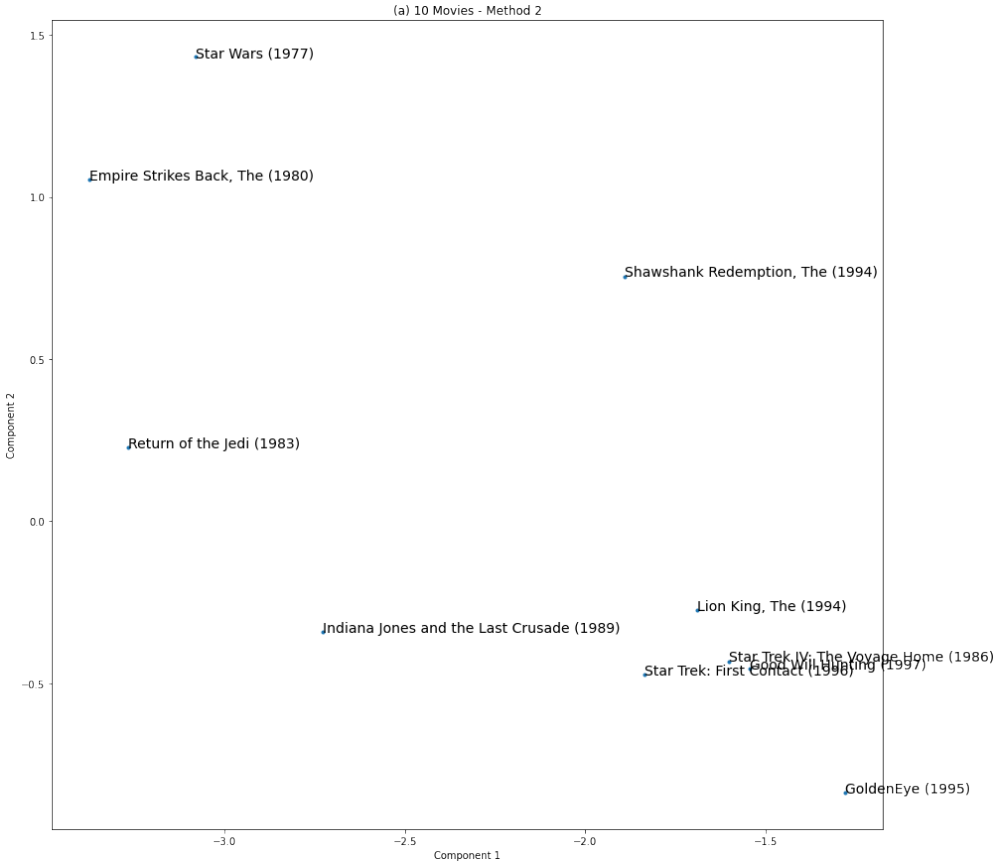
Our method 2 was using the same matrix factorization method from Homework 5 with the incorporation of bias terms a and b for each user and movie in order to model global tendencies of the various users and movies. Again, this method was in order to find the matrices U and V (such that $Y \simeq U^T$). With $k = 20$, we kept the same parameters and and stopping criteria as used in the homework in order to allow for proper comparisons. Same as the previous method, with more time, we would have performed random / grid searches to further optimize these aspects. The MovieLens dataset was split into train.csv and test.csv as given, and this method was run on train.csv with its performance assessed on test.csv using the mean regularized-squared error. To note, the SGD implementation from method 1 was updated such that we need to be able to calculate the gradients of the bias terms as follows (not including learning rate):

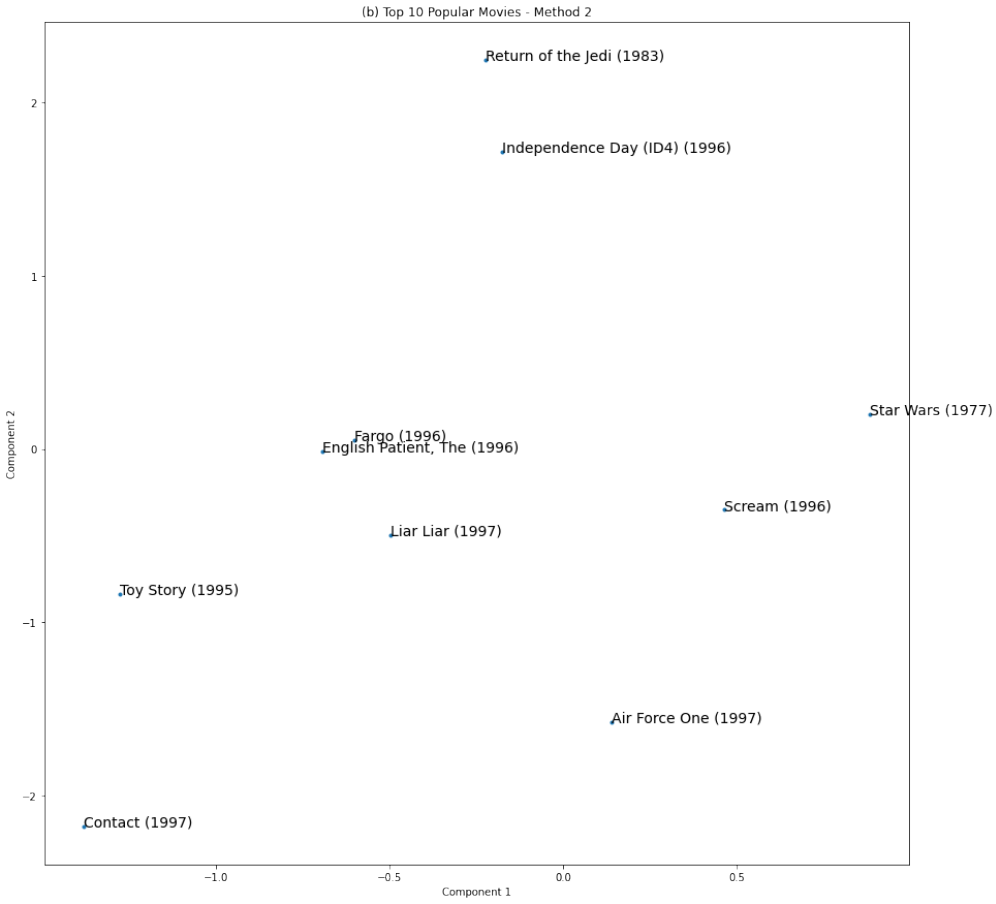
$$\delta_{a_i} = \delta_{b_j} = -\sum_{(i,j)\in S} Y_{i,j} - u_i^T v_j - a_i - b_j$$

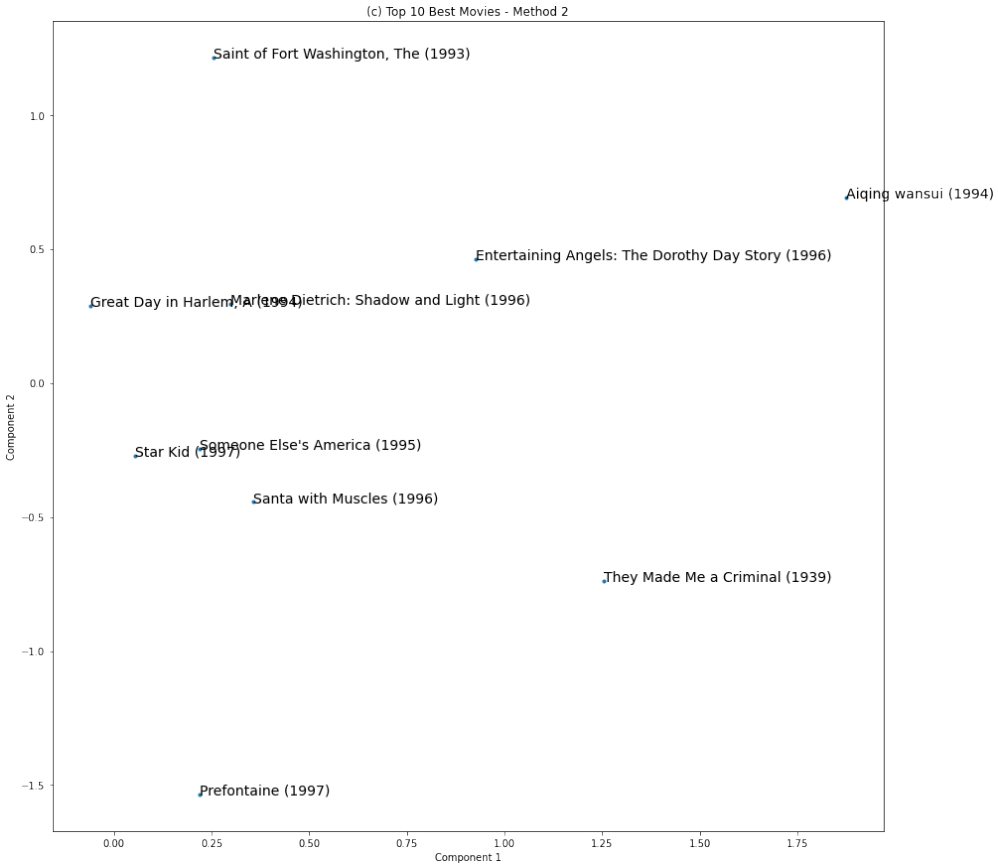
The training objective for this method is as follows:

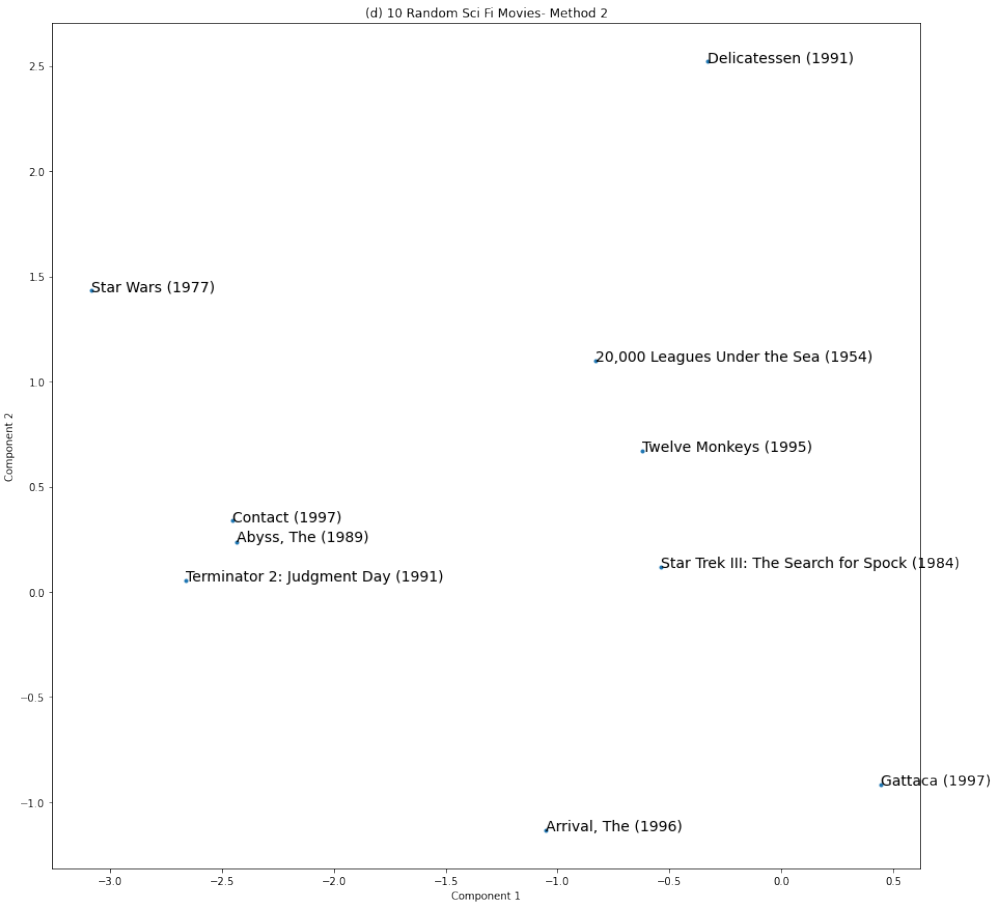
$$\underset{U,V,a,b}{\operatorname{argmin}} \frac{\lambda}{2}\left(\|U\|^2 + \|V\|^2\right) + \sum_{(i,j)\in S}\left(Y_{i,j} - \left(u_i^T v_j + a_i + b_j\right)\right)^2$$

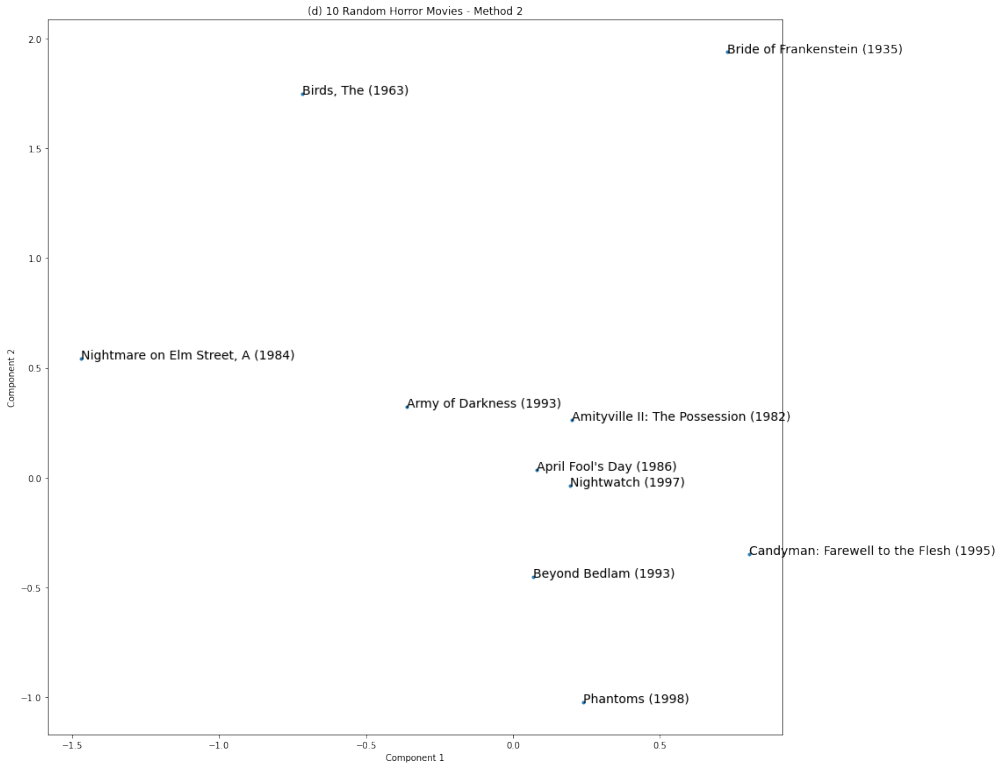
Visualizations

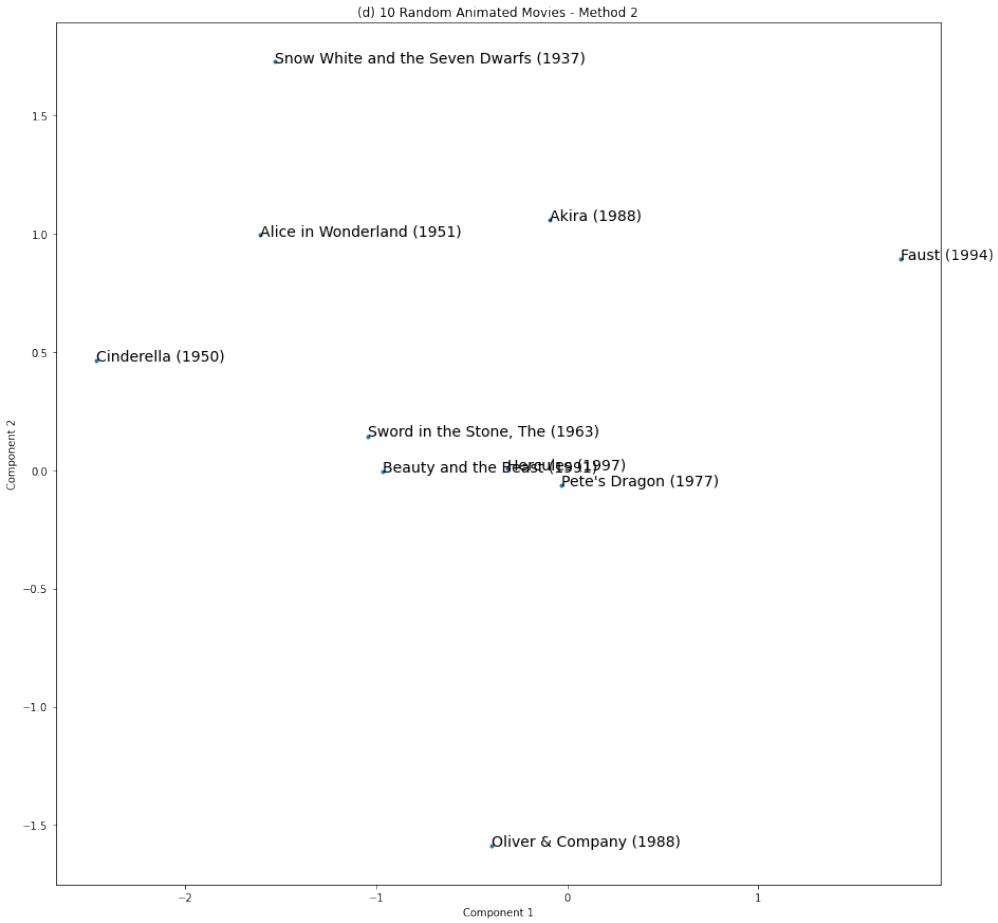



(b) Top 10 Popular Movies - Method 2
Return of the Jedi (1983)
Independence Day (ID4) (1996)
Star Wars (1977)
Fargo (1996)
English Patient, The (1996)
Scream (1996)
Liar Liar (1997)
Toy Story (1995)
Air Force One (1997)
Contact (1997)
Component 1
Component 2
2
1
0
−1
−2
−1.0
−0.5
0.0
0.5


(c) Top 10 Best Movies - Method 2
Saint of Fort Washington, The (1993)
Aiqing wansui (1994)
Entertaining Angels: The Dorothy Day Story (1996)
Great Day in Harlem, A (1994)
Marlene Dietrich: Shadow and Light (1996)
Someone Else's America (1995)
Star Kid (1997)
Santa with Muscles (1996)
They Made Me a Criminal (1939)
Prefontaine (1997)
Component 1
Component 2
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75


(d) 10 Random Sci Fi Movies- Method 2
Delicatessen (1991)
Star Wars (1977)
20,000 Leagues Under the Sea (1954)
Twelve Monkeys (1995)
Contact (1997)
Abyss, The (1989)
Star Trek III: The Search for Spock (1984)
Terminator 2: Judgment Day (1991)
Gattaca (1997)
Arrival, The (1996)
Component 1
Component 2


(d) 10 Random Horror Movies - Method 2
Bride of Frankenstein (1935)
Birds, The (1963)
Nightmare on Elm Street, A (1984)
Army of Darkness (1993)
Amityville II: The Possession (1982)
April Fool's Day (1986)
Nightwatch (1997)
Candyman: Farewell to the Flesh (1995)
Beyond Bedlam (1993)
Phantoms (1998)
Component 1
Component 2

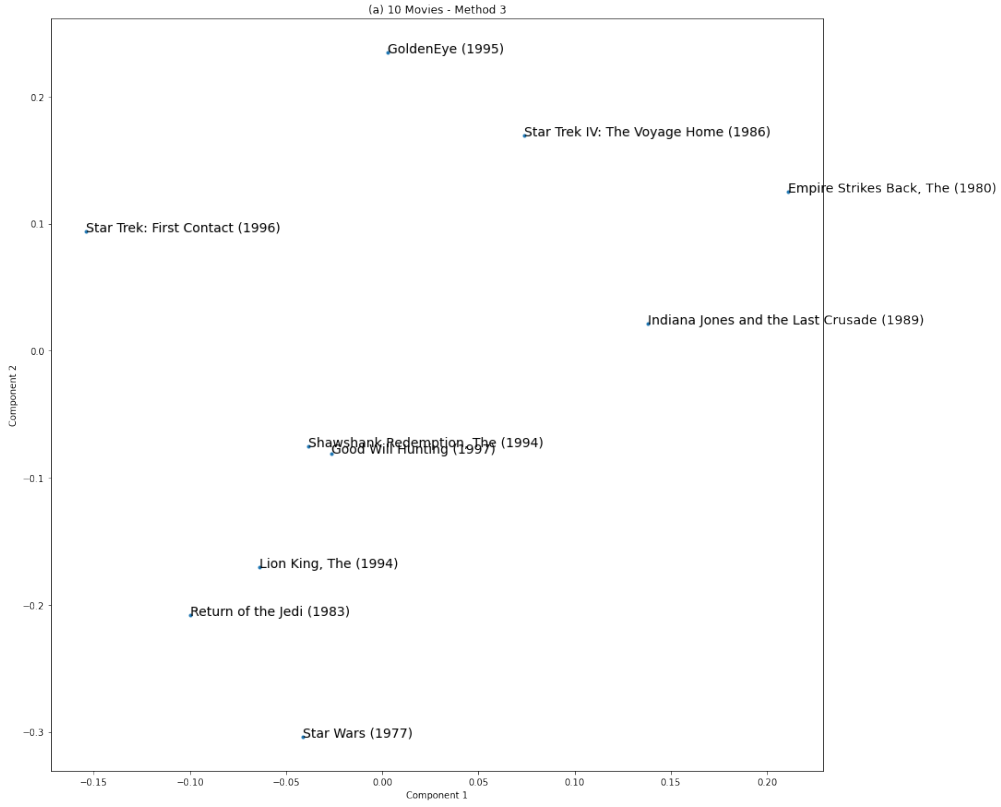


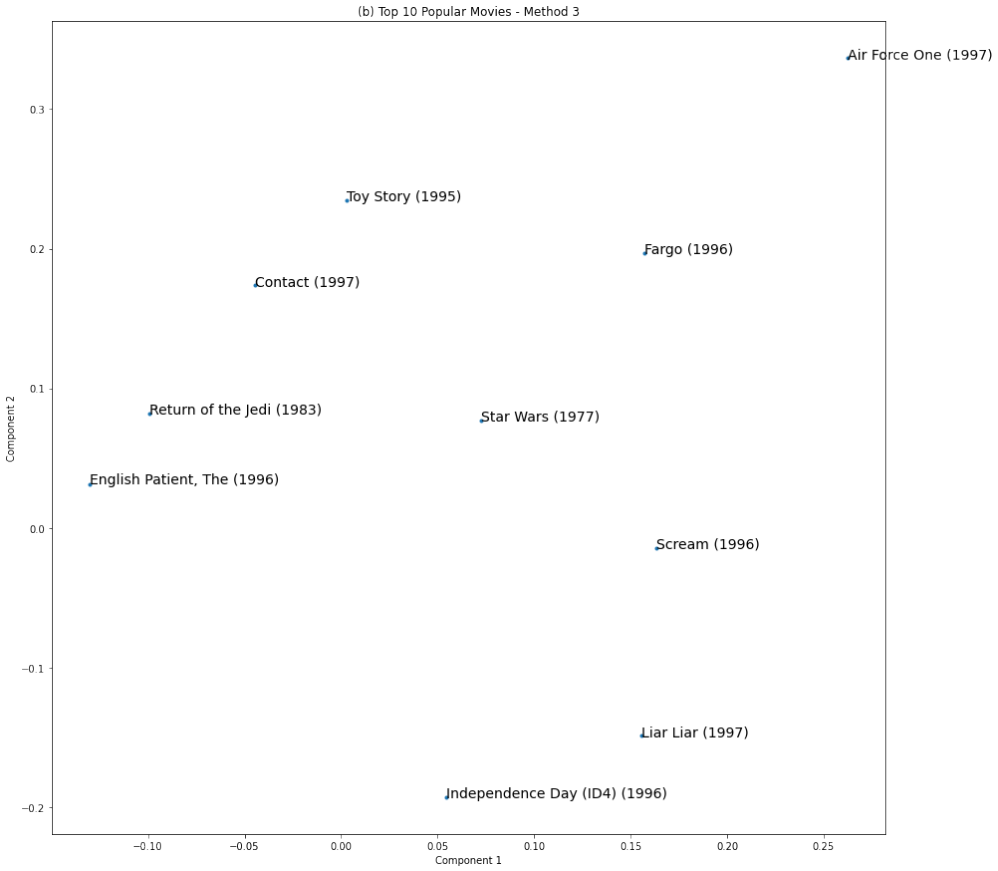
Method 3 - Off-The-Shelf (Surprise)

Description

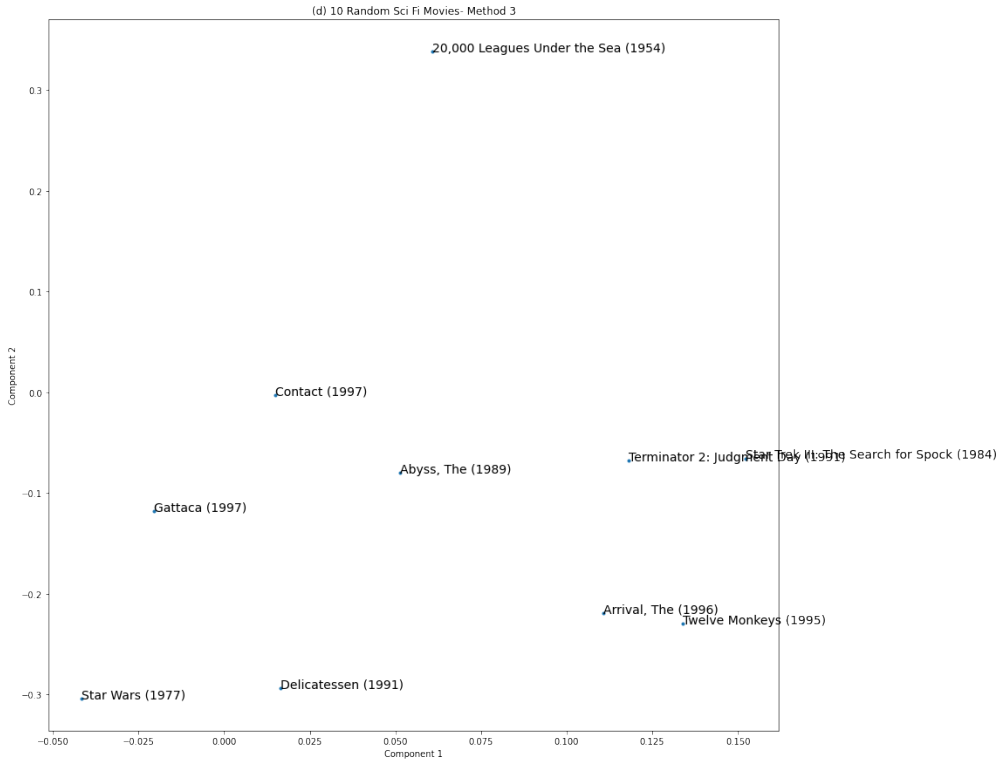
Our method 3 was using the surprise package in python. We used Surprise SVD on the training dataset to create the visualizations. After fitting the SVD, we made predictions on the test dataset to get the mean squared error to compare to the other methods. We additionally tried the cross validation function from the surprise package on the entire dataset.

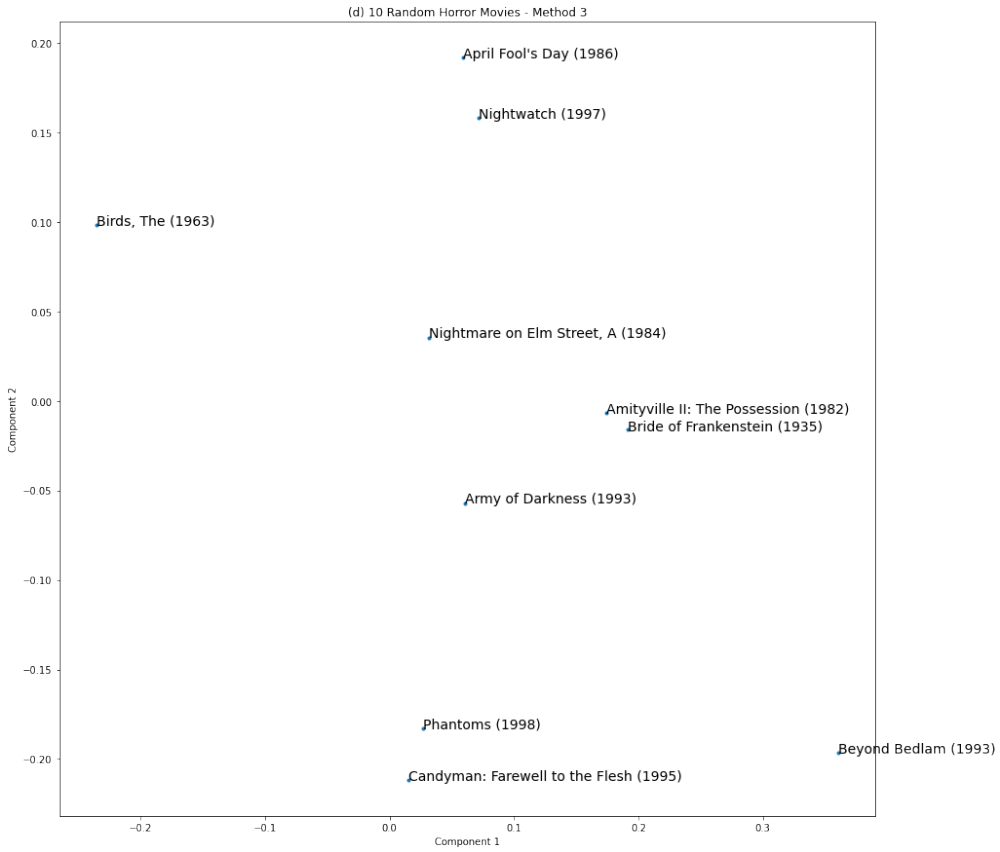
Visualizations

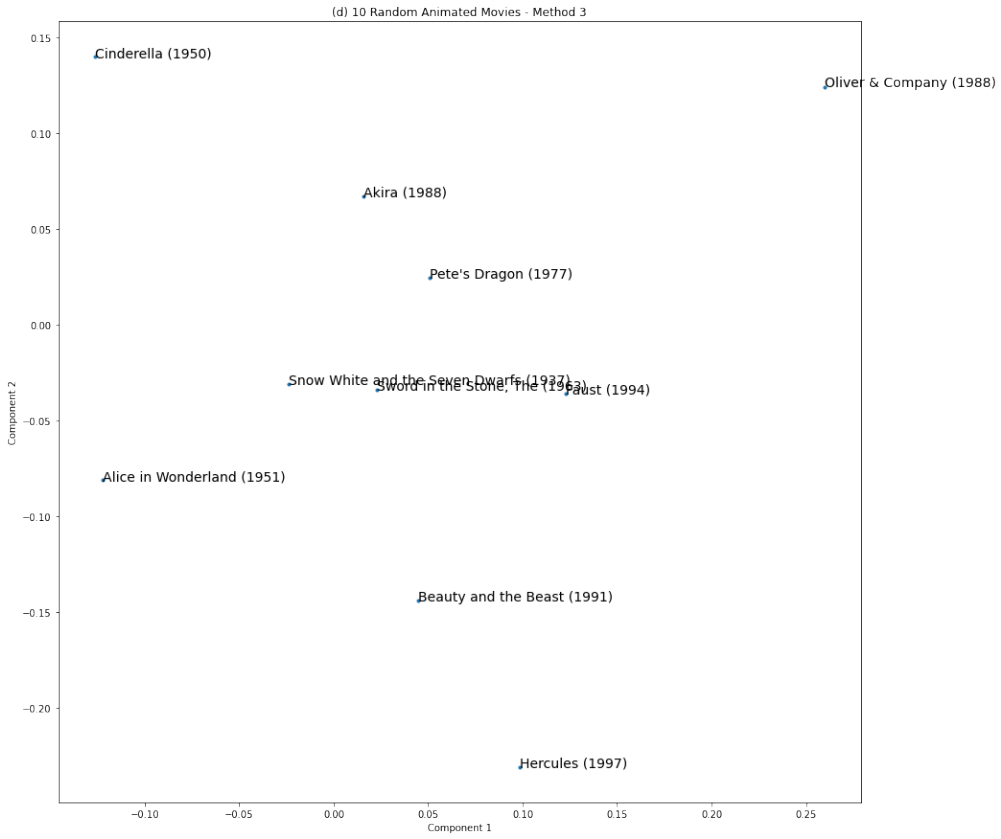



(b) Top 10 Popular Movies - Method 3
Air Force One (1997)
Toy Story (1995)
Fargo (1996)
Contact (1997)
Return of the Jedi (1983)
Star Wars (1977)
English Patient, The (1996)
Scream (1996)
Liar Liar (1997)
Independence Day (ID4) (1996)
Component 1
Component 2
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
0.25


(c) Top 10 Best Movies - Method 3
Saint of Fort Washington, The (1993)
They Made Me a Criminal (1939)
Entertaining Angels: The Dorothy Day Story (1996)
Aiqing wansui (1994)
Santa with Muscles (1996)
Great Day in Harlem, A (1994)
Star Kid (1997)
Marlene Dietrich: Shadow and Light (1996)
Prefontaine (1997)
Someone Else's America (1995)
Component 1
Component 2
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.075
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100


(d) 10 Random Sci Fi Movies- Method 3
20,000 Leagues Under the Sea (1954)
Contact (1997)
Abyss, The (1989)
Terminator 2: Judgment Day (1991)
Star Trek II: The Search for Spock (1984)
Gattaca (1997)
Arrival, The (1996)
Twelve Monkeys (1995)
Delicatessen (1991)
Star Wars (1977)
Component 1
Component 2


(d) 10 Random Horror Movies - Method 3
April Fool's Day (1986)
Nightwatch (1997)
Birds, The (1963)
Nightmare on Elm Street, A (1984)
Amityville II: The Possession (1982)
Bride of Frankenstein (1935)
Army of Darkness (1993)
Phantoms (1998)
Beyond Bedlam (1993)
Candyman: Farewell to the Flesh (1995)
Component 1
Component 2



Comparison and Analysis of Methods

The RME of method 1 on the test dataset was 0.638. The RME of method 2 on the test dataset was 0.649. The RME of the off-the-shelft method on the test dataset was 0.8566, much worse than the other methods. Using the cross validation method on the entire dataset for both 5 and 10 folds had even worse mean RME values than using the fit and prediction accuracy method. The off-the-shelf method performs straightforward stochastic gradient descent to minimize regularized squared error with bias so is very similar to the other methods used; this doesn't seem to explain why it performs so much worse than the other methods, so it is likely just the parameters that make up this difference. Moreover, these differences in performance can generally be seen in the visualizations such that methods 1 and 2 seem to have stronger (and similar) trends while method 3 has weaker trends. Another difference to highlight is that the presence of the bias terms seems to be a factor in the difference between the frames in which we are viewing these movie visualizations. In particular, taking into account the axis ranges for both components, the latent space plots for method 1 are generally in smaller windows that in method 2. This difference seemingly has an impact on certain trends. These differences are further discussed below.

Most Popular vs. Best

For the most popular movies, we would expect to see some sort of pattern given that these movies give us more data to work with (more reviews). For the best movies, we were unsure whether to expect strong

trends due to the low number of reviews for the best movies by highest average ratings. With the most popular movies, particularly with method 1, the movies generally lie on a diagonal line from bottom left to upper right which is an interesting trend. This would make us think that movies in the upper left and bottom right regions would not be very popular, generally. This trend is also somewhat present using method 2. However, with method 3, this trend does not hold as well. With the best movies, we do not really see any strong trends with any of the methods. This makes sense because the movies with the highest average ratings likely have very few reviews (as discussed in basic visualizations). Perhaps if we would have used some review count threshold for the "best" movies, then we would see a stronger relationship between the plotted movies. Overall, it makes sense that methods 1 and 2 exhibit similar trends, particularly with the most popular movies, considering their similar performance on the test set. Furthermore, since method 3 did not perform as well as the other methods, it makes sense that its trends were not as strong. Thus, our expectations were mostly met.

Sci-Fi vs. Horror vs. Animated
In the visualizations, we can see differences in where the movies are concentrated for the three differ genres. The locations of the different genres on the graphs do different between methods, but we can see that, for example, in method 3, we have sci-fi movies clustered on the lower half of the graph (only 1 of the 10 movies above y=0), horror movies clustered on the right half of the graph (only 1 of the 10 movies below x=0), and animated movies not having a particular cluster.

Method 1 vs. Method 2 vs. Method 3
In the visualizations, we can see that the different methods produce some similarities and differences in principal components by where the movies and genres are located. For example, while in method 3 we saw sci-fi movies clustered on the lower half while in method 2 we see them clustered on the left half of the graph (below x=0), but for method 1 we again see most of the movies on the lower half (7 movies below and 2 close to x=0). Looking at specific movies, we see that in method 2, we see the two Star Trek movies close to each other and the three Star Wars movies fairly close to each other, but for the other two methods these movies are not grouped in this way. As another example, with the Sci-Fi genre, we see that Delicatessen (1991) is an outlier in methods 1 and 2, but not in method 3. This same trend can be seen with Oliver and Company (1988) for the animation genre.

In conclusion, we are satisfied with our results. With more time, we would done more to optimize each method and perhaps performed further comparisons between the methods. We also would have experimented more with the off-the-shelf method(s). According to our calculation, method 1 performed slightly better than method 2 while method 3 did not perform that well compared to the other two methods. We wanted to keep our visualizations simple and straightforward in order to learn the most about what our methods were producing for U and V (after projections, rescaling, etc.). Overall, it seemed like we reached a 2-dimensional space that is representative of the 943 users and 1682 movies, considering some of the trends we saw.