Machine Learning & Data Mining
Miniproject 3

Caltech CS/CNS/EE 155
Released March 3^{rd}, 2017

---

1 Introduction

- **Group members**
  Vaibhav Anand
  Nikhil Gupta
  Michael Hashe
- **Team name**
  The Breakfast Club
- **GitHub link**
  https://github.com/nkgupta1/CS155-Visualization
- **Division of labour**
  Vaibhav Anand: Matrix Factorization Visualizations, Report
  Nikhil Gupta: Matrix Factorization Algorithm, Report
  Michael Hashe: Basic Visualizations, Report

2 Basic Visualizations

Justify your choice of visualization method

For each of the basic visualizations, we displayed data in (normalized) histogram format. This method was suggested by the project guidelines, and is also a very clear way of comparing frequency of ratings. This method is well-suited to visualizing data involving counts (i.e., no error bars, variation, etc.), which is what we are asked to do in this section. For Part 3, we provide two graphs; the top 10 highest rated movies (all of which have only been rated 5), and the top 10 highest rated movies with at least 50 ratings (these also happen to be the top 10 highest rated movies with at least 10 ratings, so the cutoff isn't too important).

What did you observe?

For Part 1, in which we visualized all ratings, we observed that 3's and 4's were present more frequently than would be expected in a uniform rating scheme (~ 0.2 frequency for each), 5's were present exactly as frequently as expected, and 1's and 2's were underrepresented (Fig. 1).

For Part 2, in which we visualized the 10 most popular movies as determined from the number of rankings (A.1, Fig. 2), we found that the most frequent ratings were 4's and 5's, both of which increased in frequency compared to the total dataset. The frequency of 1's, 2's, and 3's decreased.

For Part 3, in which we visualized the 10 highest rated movies (A.2, Fig. 3), we found (somewhat unsurprisingly) that all ratings were 5's. When we only considered movies with sufficiently many ratings (A.3, Fig. 4), we found that the most common rating was still 5, with 4's also highly represented. Lower ratings were uncommon, with around 12% of total ratings being 3's or below.

For Part 4, in which we considered ratings for particular genres, we examined Fantasy (Fig. 5), Sci-Fi (Fig. 6), and War movies (Fig. 7). Sci-Fi movies followed the same general trend as overall ratings, while both Fantasy and War movies tended to be skewed towards high ratings but otherwise similar.

Did the results match what you would expect to see?

For Part 1, the results matched our expectations. We would expect that better movies would be watched more, which would lead to higher ratings.

For Part 2, the results were close to our expectations. We personally would have expected the most popular movies to have an overwhelming majority of 5's; while the most popular movies are significantly more highly rated than the general dataset, their most common rating is still a 4.

For Part 3, the results matched our expectations, both for the top 10 highest rated movies and the top 10 highest rated movies with at least 50 ratings. In the first case, we expected that there would be movies with very few ratings, and of those several would only have 1 or 2 ratings of 5 (as it turns out, there were exactly 10 of them). In the second case, it makes sense that the most highly-rated movies would have ratings near to 5 (in fact, the top ratings were closer to 4.5); as such, we would expect a sizable majority of ratings to be 5's, with very few low ratings.

For Part 4, the results matched our expectation. Given that these are 3 fairly large (and overlapping) genres, we expected that they would match both the overall ratings frequencies and the frequencies of each other. While Fantasy and War movies skewed toward higher ratings than Sci-Fi, the general patterns were very similar.

How do the ratings of the best movies compare to those of those of the most popular movies?

As mentioned above, the ratings of the most popular movies tend to be somewhat similar to the ratings of the overall dataset, while the ratings of the highest-rated movies ($\geq$ 50 ratings) are heavily skewed towards 5's. we would have expected the ratings of these two graphs would be more similar. To explain this difference, it appears that the more popular movies are aimed towards a more general audience (i.e. Toy Story, Independence Day). In contrast, many of the most highly-rated movies are somewhat older and (in my opinion) more culturally significant (i.e. Casablanca, 12 Angry Men); if we assume that user ratings are fairly recent, then it is possible that their ratings for these movies are boosted from a targeted audience that is specifically searching for them, whereas ratings for the most popular movies are from a more representative sample of the general public.

How do the ratings of the three genres you chose compare to one another?

They are very similar. This is not unexpected, as the Fantasy, Sci-Fi, and War genres overlap significantly (i.e., Star Wars). Further, all of these genres appeal to a fairly similar demographic, so we would expect the users ratings these movies to have similar rating patterns.

3 Matrix Factorization Algorithm

What parameters did you adjust and how?

We adjusted:

- the learning rate η
- the regularization strength λ
- the stopping parameter ϵ

For the ways we adjusted η, λ, please see Appendix C. How we adjusted ϵ is described in the next subsection*.

Justify your choices for the parameters and stopping criteria

We grid searched for values of the η, λ that would provide the best test error. We broke up the data into a training set including $\frac{2}{3}$ of the data and a test set containing the remaining $\frac{1}{3}$ of the data. Please see Appendix C for results. In addition, we tried various stopping criteria and ended up settling on $\epsilon = 0.003$ as the one that gave the best test error. In order to determine this, we tried different orders of magnitude (0.0001, 0.001, 0.01, 0.1) for the value and after finding the best one (0.001), increased the resolution of the search (0.001, 0.002, 0.003, 0.004, 0.005) ending up settling on a specific value (0.003). One important note is that we chose our parameters such that the model would generalize best. In other words, we chose parameters to decrease out of sample error. However, this came at a cost of increasing in sample error but as this is a predictive model, we thought that out of sample error is more important than in sample error.

Did you make any other significant modifications or additions

We tried various methods of training:

- excluding regularization error in calculating stopping criterion
- including regularization error in calculating stopping criterion
- removing the mean from the ratings
- adding a bias term for individuals and movies and training on that

Each of these changes progressively made the out of sample error better.

4 Matrix Factorization Visualization

What did you observe?

We observed a significant correlation in the rating of a movie and one of the 2 dimensions of $\widetilde{V}$ (Appendix D). Although there was no discernible correlation between movie genre and either of the dimensions, we

noticed that movies belonging from the same series (like sequels of Star Trek movies) appeared to be clustered closer together compared to random noise (Appendix D). We also attempted to find correlations between popularity and dating of the movies and the two dimensions. The release date of a movie appeared to have none (Appendix D). Note that for many of these observations, unless otherwise specified, we only plotted coordinates for movies with a minimum number of ratings (usually 50). We found that this helped see potential correlations more clearly because the ones with more ratings would express a stronger correlation.

How do the ratings of the best movies compare to those of the most popular movies

Popularity appeared to have a slight correlation in the same dimension and direction as rating in $\widetilde{V}$ (Appendix D). We believe this is because popularity is inherently correlated with the rating of a movie, meaning that more popular movies tend to have higher ratings, and since the correlation is stronger with $\widetilde{V}$ and rating, we believe the popularity correlation is a side effect of that.

How do the ratings of the three genres you chose compare to one another

As shown in Figure (Appendix D), there is no difference in the values of the two dimensions $\widetilde{V}$ and the three genres. Overall, the distribution of ratings for the three genres is similar and is explained in more depth and shown in section (2).

What was expected and what was surprising from the visualizations?

We expected to see a stronger (or any) correlation between the 2 dimensions of $\widetilde{V}$ and the genres, as shown in the examples in class. Even after implementing bias and bias regularization in the matrix factorization and grid-searching to find the optimal step size and regularizations, the visualization of $\widetilde{V}$ after performing SVD did not result in any meaningful correlations besides the average rating of movies, which is rather obvious with ML.

Any other comparisons/observations

While there may be correlations between features in the movies given by $\widetilde{V}$, they go beyond genres and dating, and have esoteric meaning. More detailed observations on individual genres can be seen in (Appendix D.1).

Also, before we implemented bias and bias regularization in the matrix factorization, the values of $\widetilde{V}$ for the movies were in the positive quadrant, the correlation between popularity and rating appeared to be stronger, and E_{out} was higher. However, after bias, the data became more centered in both dimensions.

5 Conclusion

Briefly summarize your main observations

We found that we could optimize the parameters of our model consistently and effectively to minimize E_{out}. We also found that adding bias improved our E_{out}. However, given the amount of data and how much true correlation we would expect to exist between users and movies, we believe the produced models were sufficient for visualization analysis. Our model achieved an MSE of less than 0.5, which implies that it can predict movie ratings for a user within a point.

Did your visualizations help you to better understand the MovieLens dataset?

The observations from the basic visualization part were expected, but they also helped us in understanding the results from the matrix factorization visualization. For example, we noticed that popularity and ratings were highly correlated, which explained why we saw, at least initially, a strong popularity correlation between popularity and the coordinates in $\widetilde{V}$.

There was no observable trend we could find from the matrix factorization visualization and potential features such as genres, time, sequels, popularity, etc. Instead, the prediction of rating was more dependent on the average rating for a movie, which we would expect at minimum. There may exist deeper patterns that were found in the movies but are indiscernible from the visualizations, but, overall, we did not see the amount of correlation that we had expected.

Appendices

A Basic Visualizations Data

A.1 Most Popular Movies

```
Star Wars (1977)
Contact (1997)
Fargo (1996)
Return of the Jedi (1983)
Liar Liar (1997)
"English Patient, The (1996)"
Scream (1996)
Toy Story (1995)
Air Force One (1997)
Independence Day (ID4) (1996)
```

A.2 Highest Ratings (All Movies)

```
"Great Day in Harlem, A (1994)"
They Made Me a Criminal (1939)
Prefontaine (1997)
Marlene Dietrich:  Shadow and Light (1996)
Star Kid (1997)
"Saint of Fort Washington, The (1993)"
Santa with Muscles (1996)
Aiqing wansui (1994)
Someone Else's America (1995)
Entertaining Angels:  The Dorothy Day Story (1996)
```

A.3 Highest Ratings ($\geq$50 Ratings)

```
"Close Shave, A (1995)"
Schindler's List (1993)
"Wrong Trousers, The (1993)"
Casablanca (1942)
Wallace & Gromit:  The Best of Aardman Animation (1996)
"Shawshank Redemption, The (1994)"
Rear Window (1954)
"Usual Suspects, The (1995)"
Star Wars (1977)
12 Angry Men (1957)
```

B Basic Visualizations

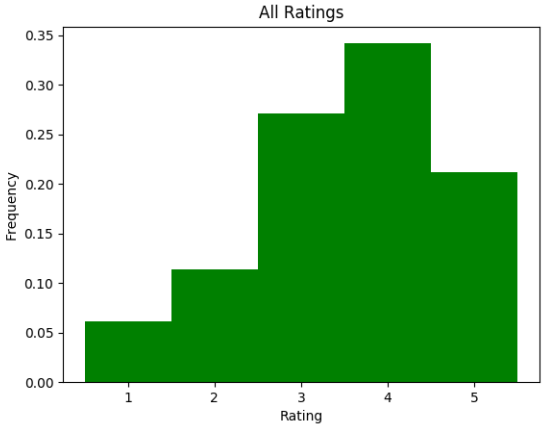

Figure 1: Histogram of all ratings in dataset.

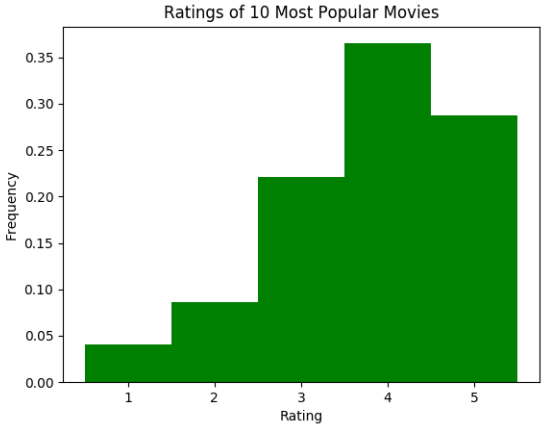

Figure 2: Histogram of 10 most popular movies.

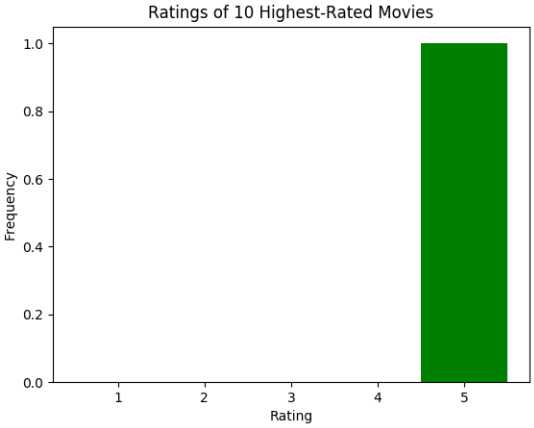

Figure 3: Histogram of 10 highest-rated movies.

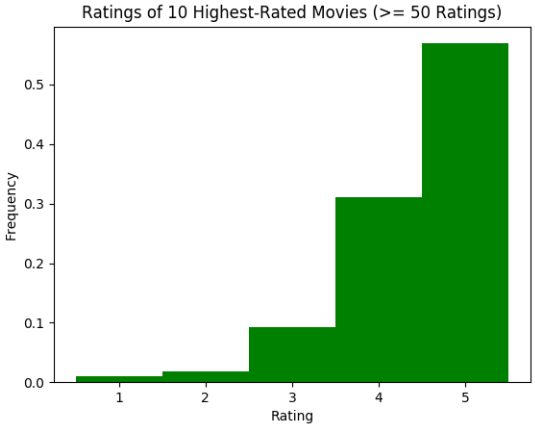

Figure 4: Histogram of 10 highest-rated movies with over 50 ratings.

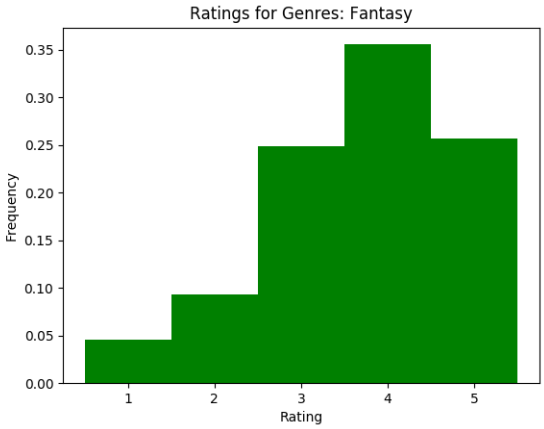

Figure 5: Histogram of ratings for fantasy movies.

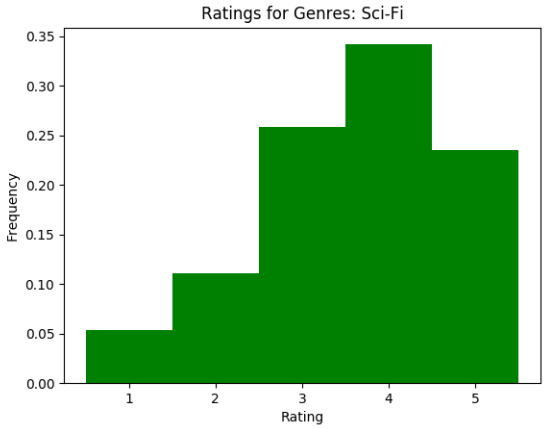

Figure 6: Histogram of ratings for sci-fi movies.

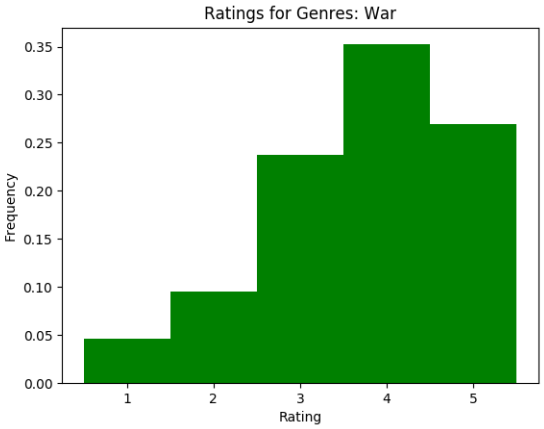

Figure 7: Histogram of ratings for war movies.

C Grid Search Results

C.1 error without regularization in training (ordered by out of sample error)

first $\frac{2}{3}$ training, last $\frac{1}{3}$ test:

```
errOut = 0.643741, reg = 0.10000, eta = 0.0100, errIn = 0.267795
errOut = 0.651725, reg = 0.10000, eta = 0.0200, errIn = 0.252123
errOut = 0.661422, reg = 0.10000, eta = 0.0050, errIn = 0.286361
errOut = 0.666840, reg = 0.10000, eta = 0.0090, errIn = 0.269846
errOut = 0.670787, reg = 0.10000, eta = 0.0100, errIn = 0.250795
errOut = 0.676144, reg = 0.08000, eta = 0.0100, errIn = 0.220056
errOut = 0.676604, reg = 0.11000, eta = 0.0100, errIn = 0.283961
errOut = 0.677733, reg = 0.18000, eta = 0.0100, errIn = 0.387222
errOut = 0.678625, reg = 0.12000, eta = 0.0100, errIn = 0.301132
errOut = 0.679735, reg = 0.10000, eta = 0.0100, errIn = 0.281514
errOut = 0.682994, reg = 0.10000, eta = 0.0100, errIn = 0.262050
errOut = 0.683584, reg = 0.10000, eta = 0.0500, errIn = 0.284322
errOut = 0.683747, reg = 0.10000, eta = 0.0100, errIn = 0.259213
errOut = 0.686960, reg = 0.09000, eta = 0.0100, errIn = 0.261056
errOut = 0.692979, reg = 0.10000, eta = 0.0100, errIn = 0.258826
errOut = 0.714738, reg = 0.10000, eta = 0.0500, errIn = 0.284227
errOut = 0.731857, reg = 0.15000, eta = 0.0100, errIn = 0.362397
errOut = 0.751690, reg = 0.10000, eta = 0.1000, errIn = 0.388069
errOut = 0.752882, reg = 0.30000, eta = 0.0100, errIn = 0.468818
errOut = 0.753859, reg = 0.20000, eta = 0.0100, errIn = 0.420194
errOut = 0.761073, reg = 0.10000, eta = 0.1000, errIn = 0.394411
errOut = 0.821591, reg = 0.40000, eta = 0.0100, errIn = 0.516962
errOut = 0.852405, reg = 0.01000, eta = 0.0100, errIn = 0.152351
errOut = 0.854635, reg = 0.50000, eta = 0.0100, errIn = 0.560920
errOut = 0.918906, reg = 0.60000, eta = 0.0100, errIn = 0.620285
errOut = 0.925896, reg = 0.00010, eta = 0.0100, errIn = 0.142089
errOut = 0.931411, reg = 0.70000, eta = 0.0100, errIn = 0.681322
errOut = 0.935970, reg = 0.00100, eta = 0.0100, errIn = 0.148187
errOut = 0.943312, reg = 0.00000, eta = 0.0100, errIn = 0.142968
errOut = 1.066705, reg = 0.80000, eta = 0.0100, errIn = 0.769256
errOut = 1.250584, reg = 1.00000, eta = 0.0100, errIn = 0.956924
```

first $\frac{1}{3}$ training, last $\frac{2}{3}$ test:

```
errOut = 0.460330, reg = 0.10000, eta = 0.0100, errIn = 0.260879
errOut = 0.461616, reg = 0.10000, eta = 0.0050, errIn = 0.274322
errOut = 0.462888, reg = 0.10000, eta = 0.0090, errIn = 0.267259
errOut = 0.467803, reg = 0.10000, eta = 0.0200, errIn = 0.270230
```

```
errOut = 0.651215, reg = 0.01000, eta = 0.0090, errIn = 0.140637
errOut = 0.659732, reg = 0.01000, eta = 0.0100, errIn = 0.145086
errOut = 0.681749, reg = 0.01000, eta = 0.0050, errIn = 0.129947
errOut = 0.687374, reg = 0.01000, eta = 0.0200, errIn = 0.151102
```

C.2 error with reg in training and mean removed

first $\frac{2}{3}$ training, last $\frac{1}{3}$ test:

```
errOut = 0.472064, reg = 0.10000, eta = 0.0050, errIn = 0.222142
errOut = 0.474038, reg = 0.10000, eta = 0.0100, errIn = 0.224702
errOut = 0.477569, reg = 0.10000, eta = 0.0010, errIn = 0.290618
errOut = 0.491969, reg = 0.10000, eta = 0.1000, errIn = 0.237443
errOut = 0.624262, reg = 1.00000, eta = 0.0050, errIn = 0.638081
errOut = 0.624280, reg = 1.00000, eta = 0.1000, errIn = 0.638152
errOut = 0.624281, reg = 1.00000, eta = 0.0100, errIn = 0.638098
errOut = 0.732767, reg = 0.01000, eta = 0.0050, errIn = 0.114247
errOut = 0.789952, reg = 0.01000, eta = 0.0100, errIn = 0.103601
errOut = 0.819212, reg = 0.01000, eta = 0.1000, errIn = 0.202040
errOut = 0.847534, reg = 0.00100, eta = 0.0050, errIn = 0.110945
errOut = 0.868467, reg = 0.00010, eta = 0.1000, errIn = 0.440801
errOut = 0.880642, reg = 0.00010, eta = 0.0050, errIn = 0.109889
errOut = 0.921586, reg = 0.00100, eta = 0.1000, errIn = 0.417915
errOut = 1.048777, reg = 0.00100, eta = 0.0100, errIn = 0.100087
errOut = 1.088209, reg = 0.00010, eta = 0.0100, errIn = 0.100609
```

C.3 error with reg in training, mean removed, bias term, stopping parameter = 0.003:

first $\frac{2}{3}$ training, last $\frac{1}{3}$ test:

```
errOut = 0.446253, reg = 0.10000, eta = 0.0100, errIn = 0.242214
errOut = 0.448719, reg = 0.10000, eta = 0.0050, errIn = 0.255420
errOut = 0.478732, reg = 0.10000, eta = 0.1000, errIn = 0.230768
errOut = 0.492864, reg = 1.00000, eta = 0.0100, errIn = 0.476065
errOut = 0.494233, reg = 1.00000, eta = 0.0050, errIn = 0.475877
errOut = 0.507058, reg = 1.00000, eta = 0.1000, errIn = 0.493825
errOut = 0.594936, reg = 0.01000, eta = 0.0050, errIn = 0.150427
errOut = 0.621251, reg = 0.01000, eta = 0.0100, errIn = 0.136140
errOut = 0.798689, reg = 0.01000, eta = 0.1000, errIn = 0.207856
```

D Matrix Factorization Visualizations

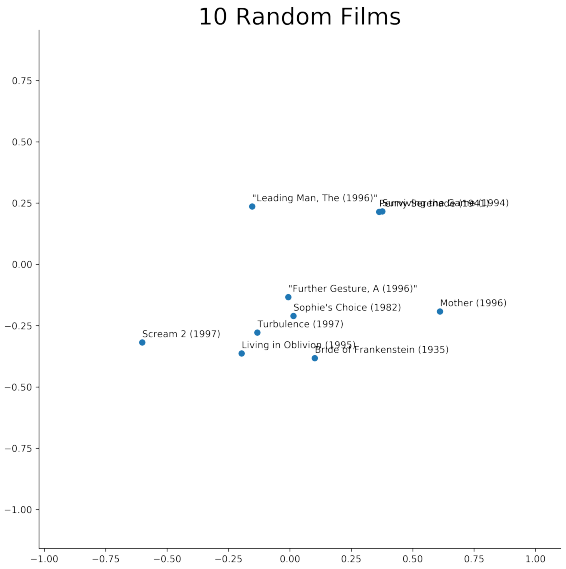

Figure 8: The coordinates of 10 randomly chosen films in the space of $\widetilde{V}$.

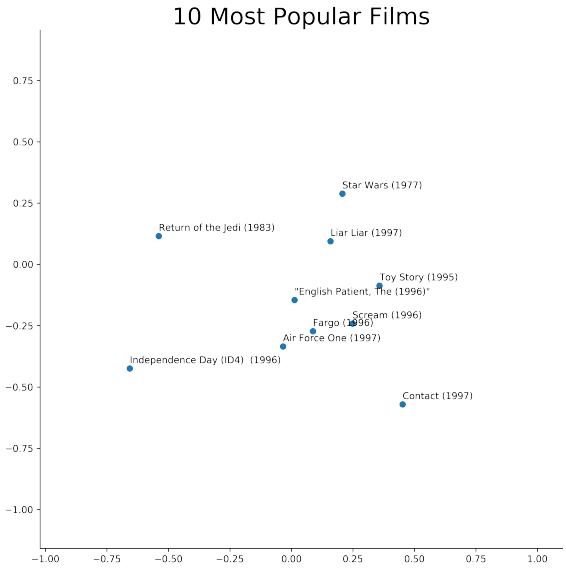

Figure 9: The 10 most popular films in the space of $\widetilde{V}$.

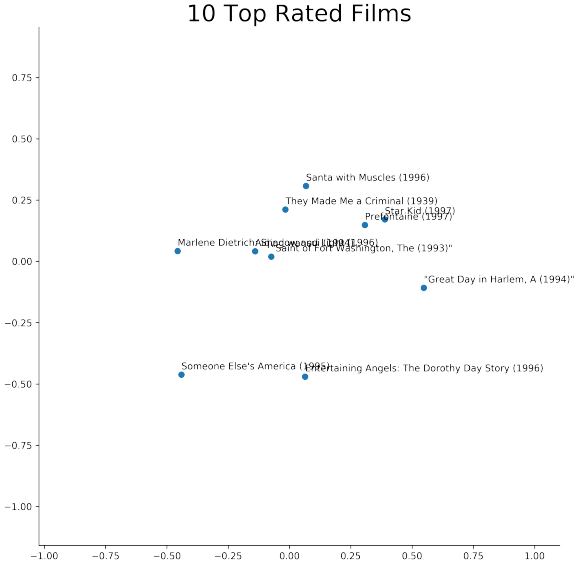

Figure 10: The 10 films with the highest average ratings and over 50 total ratings in the space of $\widetilde{V}$

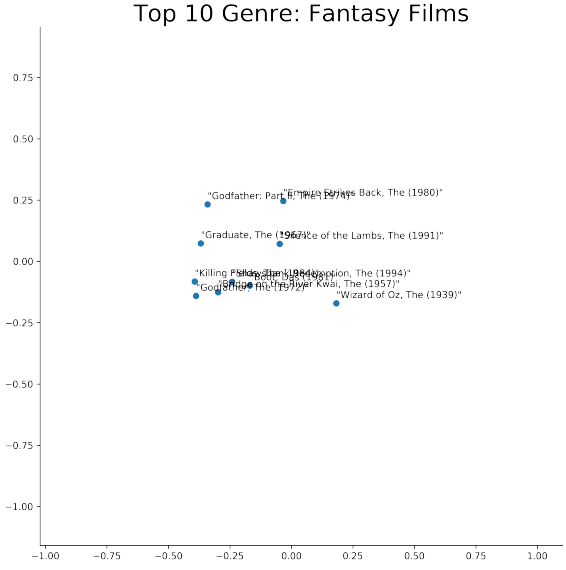

Figure 11: The top 10 most popular films in the Fantasy genre, one of the three chosen in Basic Visualizations.

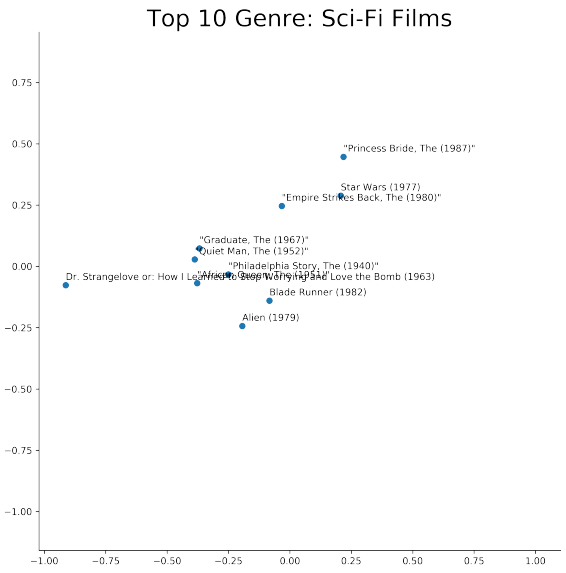

Figure 12: The top 10 most popular films in the Sci-Fi genre, one of the three chosen in Basic Visualizations.

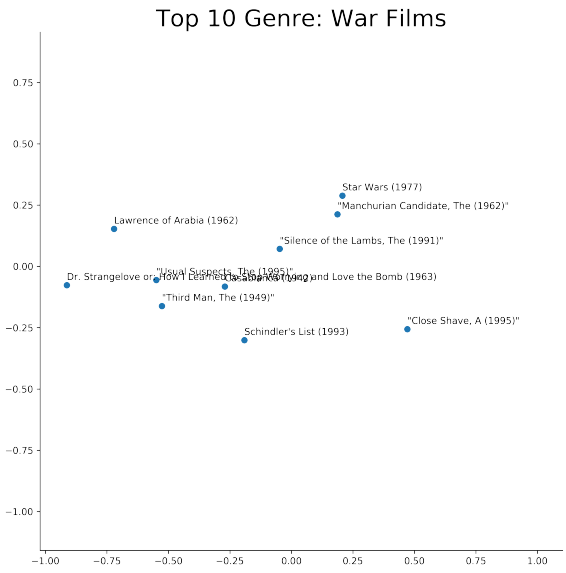

Figure 13: The top 10 most popular films in the War genre, one of the three chosen in Basic Visualizations.

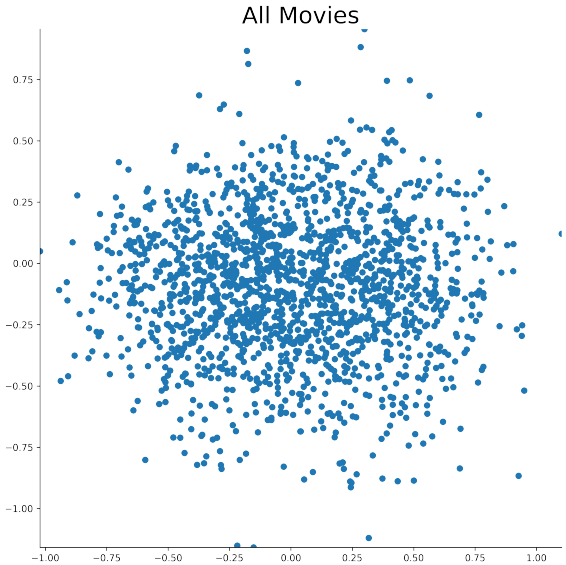

Figure 14: Coordinates of all movies in the dataset shown in the space of $\widetilde{V}$.

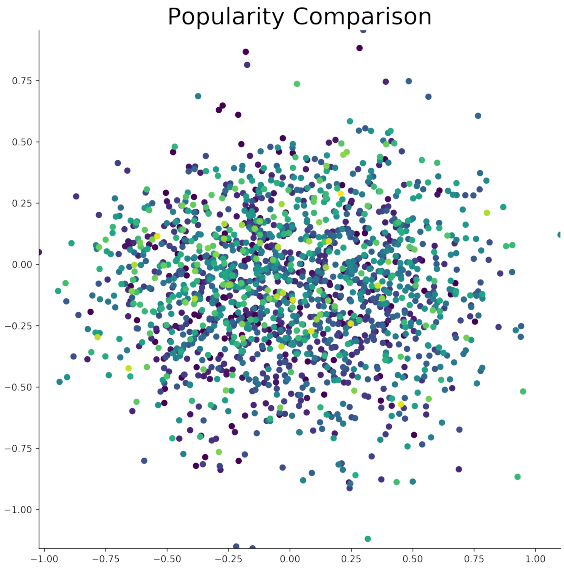

Figure 15: There is no statistically significant correlation between popularity and any of the dimensions of $\widetilde{V}$.

Figure 16: The only real correlation we found is seen in this plot, where rating is correlated with the horizontal dimensions of $\widetilde{V}$.

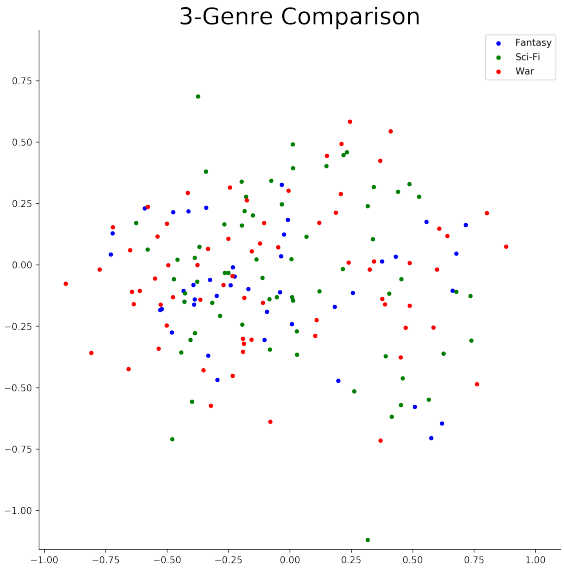

Figure 17: Comparison of coordinates of all movies from the three genres, plotted separately in $\widetilde{V}$.

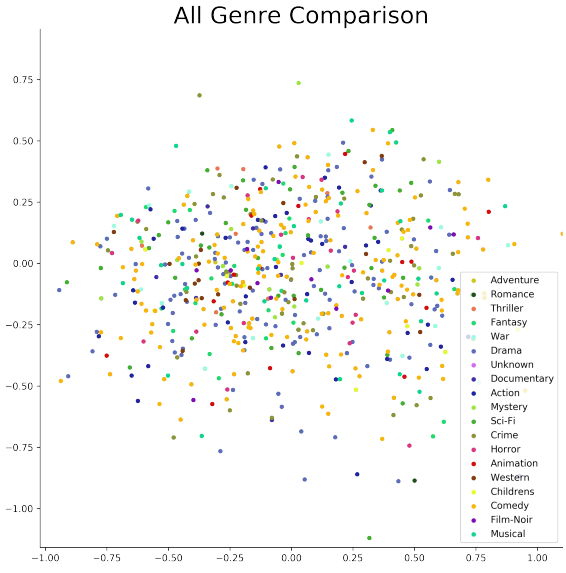

Figure 18: Comparison of all genres. Note that due to genre-overlap, genres are draw in random order shown by the descending order in the legend.

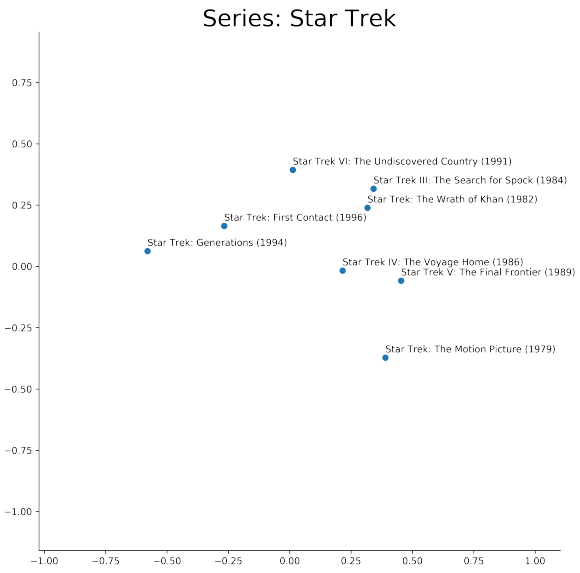

Figure 19: Comparison of an example movie series. Star Trek was chosen because it was one of the largest sequels found.

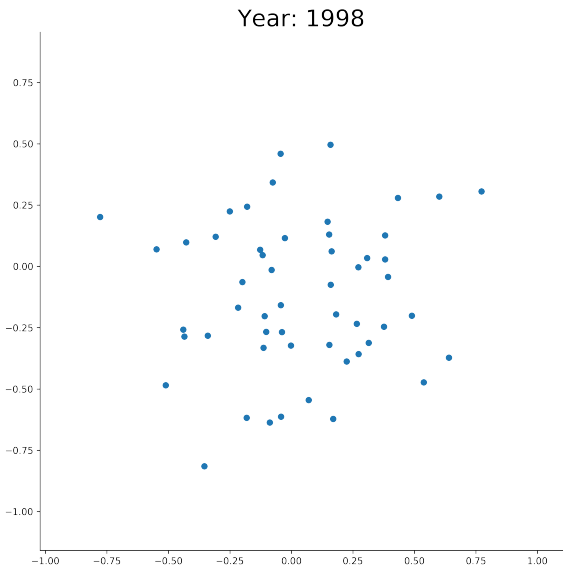

Figure 20: All movies from 1998 are plotted here. No noticeable difference from all movies noticed here either.

D.1 Genre Distributions

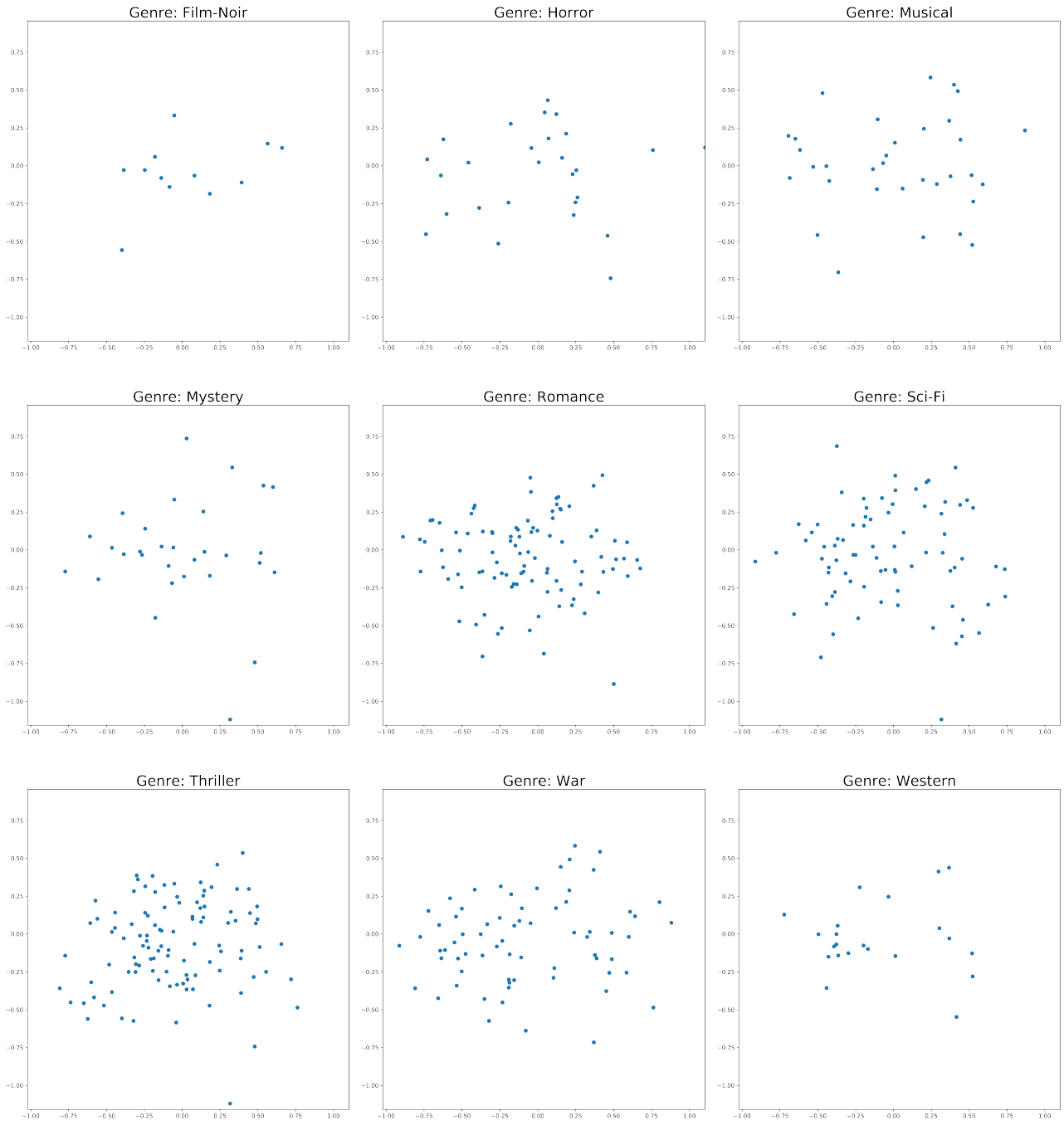

Figure 21: All the individual genres all drawn above separately. Although it might be possible that some possess some clustered in $\widetilde{V}$, nothing is significant enough to draw any conclusions.