1 Introduction [5 points]

- Group members: Sergio Esteban, Lena Wu, Asav Kumar
- Colab links:

  Basic Visualizations: https://colab.research.google.com/drive/1IcPvF˙6GLEk56klHLPnheT5Je9oNEACp?usp=sharin

  Methods 1,2,3: https://colab.research.google.com/drive/14gBQpJDud5TfAK113We3JK2KymmN34rs?usp=sharingsc
- Piazza link: https://piazza.com/class/lbv0docn6037fw/post/585
- Division of labor:
  - Lena and Asav – Matrix factorization visualizations (Methods 1,2,3), Piazza visualization post, wrote parts of report.
  - Sergio – Basic visualizations, wrote parts of report
- Packages used: Matplotlib, Surprise SVD, Numpy, Seaborn, Sklearn, Pandas

2 Basic Visualizations [20 points]

Discussion

The results are very much as one would expect. In the overall data we see what seems to be a normal distribution (Figure 1). All subsets of the dataset also exhibit normal distributions (see Figures 1—6). The average movie is scored with $\mu = 3.55$ rating and the standard deviation here was calculated as $\sigma = 1.03$. The true distribution could be normal. It's interesting to see that there is more weight to the right side of the distribution. Perhaps people are more lenient in rating movies.

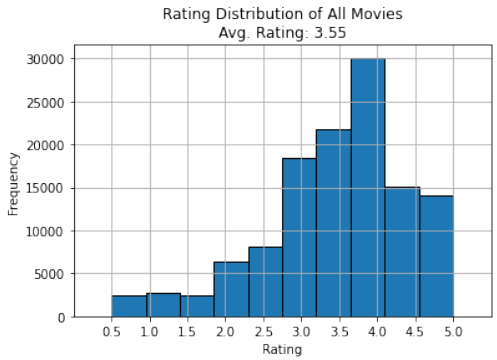

Figure 1: All ratings in the MovieLens Dataset

Now, for the top 10 most popular movies, we see a bias towards the higher end of the rating system (Figure 2). There are several possible explanations. This time, we have $\mu = 3.93$ and $\sigma = 0.96$. A possible explanation is that people may be influenced by the *bandwagon bias*—a form of bias that causes one to agree with group thinking. It is possible that individuals may think that ta movie is good from the sole reason that the majority of others rate it so highly. The same could be said about the rating data for the top 10 best movies (Figure 3). Here, we have $\mu = 4.24$ and $\sigma = 0.76$, an even higher average rating and tighter distribution! This is as expected since we filtered out the highest-rated movies from the bunch.

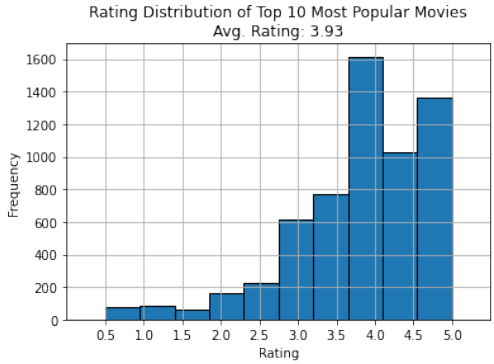

Figure 2: All ratings of the ten most popular movies (movies which have received the most ratings).

When comparing the ratings for the 10 most popular movies and the 10 best movies, it's easy to see that the average is higher on the best movies and the total number of samples is higher on the most popular movies. Of course, this is by construction since we cherry picked this data. However, we noticed that popularity of a movie does not equate to higher ratings on a movie. This is true as there is no intersection in between the set of most popular movies and the set of best movies.

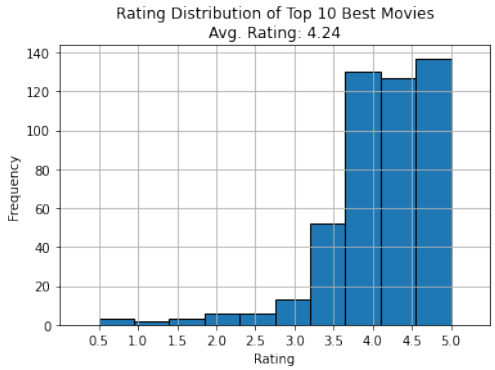

Figure 3: All ratings of the ten best movies (movies with the highest average ratings).

The most interesting subsets of data to consider were data by the genre of the movie. We chose our top three favorite movie genres: 1) Comedy, 2) Documentary, and 3) Horror (see Figures 4—6 respectively). In fact, we made additional visualizations for the Horror movie genre! Table 1 contains metrics for each movie genre.

Genre	μ	σ	No. Ratings
Comedy	3.43	1.07	39,372
Documentary	3.67	0.99	1,690
Horror	3.32	1.06	7,471

Table 1: Metrics for each movie genre.

One can see that comedy is the most popular type of movie, documentaries are the most highly rated movies, and horror movies are the lowest rated but are still popular. Additionally, documentaries have higher ratings but also smaller standard deviations. A naive look indicates that documentaries are the best kinds of movies among these three categories.

Overall, it is extremely difficult to extract meaning from these basic visualizations. Essentially, here we have two reliable metrics—ratings and number of ratings. In actuality however, the features of this movie data live in a much higher dimension. What we are seeing is only a projection onto our two metrics. This is exactly a pitfall of using this kind of visualization. We can take advantage of not only genres of the movies, but also other movie attributes such as release date of the movie, length of the movies, director of the movie, and much more.

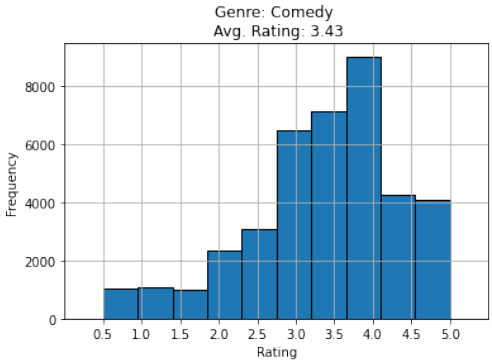

Figure 4: All ratings of movies from the genre Comedy.

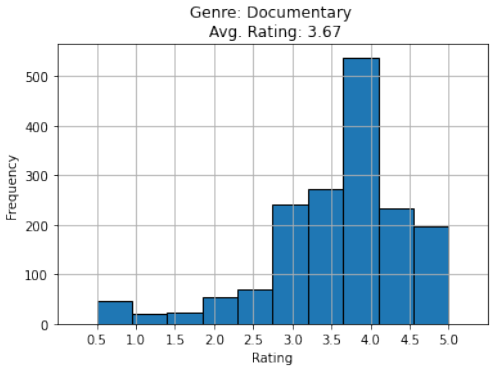

Figure 5: All ratings of movies from the genre Documentary.

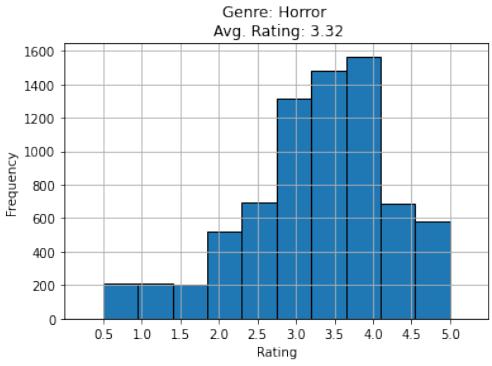

Figure 6: All ratings of movies from the genre Horror.

3 Matrix Factorization Visualizations [60 points]

Matrix Factorization Methods

Method 1: Homework 5 SVD — with Bias

We used the HW5 SVD algorithm to minimize the regularized square error so that we factorize Y:

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

This is the canonical version of SVD latent factorization where we estimate Y given a limited number of input data.

HW5 algorithm:

$$\arg\min_{U,V} \frac{1}{2}\sum_{i,j}(y_{i,j} - u_i^T v_j)^2 + \frac{\lambda}{2}(|U|_F^2 + |V|_F^2)$$

The algorithm updates with the rule:

$$u_i = u_i - n \cdot d_{ui}$$

$$v_j = v_j - n \cdot d_{vj}$$

We used the following hyperparameters on our model: For the regularization strength, we use a value of 0.0, for the stopping point, we used the default value of $\epsilon = 0.0001$, and we used a learning rate of $n = 0.03$ with 300 epochs of training.

Method 2 & 3: Surprise SVD — with & without Bias

The Surprise SVD matrix factorization is similar to the HW5 Matrix Factorization. We include bias terms b_u and b_i to represent user and movie biases. The Surprise SVD is an SVD algorithm where a prediction $\hat{r}_{ui}$ is calculated by: $\hat{r_{ui}} = b_u + b_i + v_i^T u_u$

The values for v_i and u_u are obtained from matrices U and V respectively. The objective function for this factorization method is:

$$\sum_{r_{ui} \in R_{train}} (r_{ui} - \hat{r_{ui}})^2 + \lambda(b_i^2 + b_u^2 + ||v_i||^2 + ||u_u||^2) \tag{1}$$

The SGD algorithm shares similarities with our HW5 factorization and can be expressed as:

$$b_u \rightarrow b_u + n(e_{ui} - \lambda b_u)$$
$$b_i \rightarrow b_i + n(e_{ui} - \lambda b_i)$$
$$u_u \rightarrow u_u + n(e_{ui} v_i - \lambda u_u)$$
$$v_i \rightarrow v_i + n(e_{ui} u_u - \lambda v_i)$$
$$e_{ui} = r_{ui} - \hat{r}_{ui}$$

This includes a bonus feature that allows for bias term incorporation and our model was specifically trained with 100 epochs and a regularization strength of $\lambda = 0.0$. The without-bias model should behave similarly to our HW5 implementation as it is not that different. However, the model should theoretically be able to cluster the movies better and have improved test accuracy.

Visualizations

Visualization Discussion

Test Errors:

HW5 SVD No Bias: 0.5803

Surprise SVD, Bias: 0.3754

Surprise SVD, without Bias: 0.4695

Method	E_{out}
1. HW5 SVD	0.5803
2. Surprise SVD, Bias	0.3754
3. Surprise SVD, no Bias	0.4695

Table 2: Summary of visualization plots.

Visualization Plots

Tabulated below are Figures for each type of method and movie selection.

Method	by Choice	Popular	Best	Genre
1. HW5 SVD	7	8	9	10—12
2. Surprise SVD, Bias	13	14	15	16—18
3. Surprise SVD, no Bias	19	20	21	22—24

Table 3: Summary of visualization plots.

Movies by Choice, Popular Movies, Best Movies - with vs without bias

The same movies selected by choice have been visualized quite differently by the two models. As expected, the models differ in the number of latent factors taken into account. With bias, the Surprise SVD model has both strong horizontal **and** vertical gradients with allocating entirely different positions to the same movies when compared to its without bias version. On the other hand, without bias seems to cluster the movies away from the (Feature 1, Feature 2) origin and to the top right with a strong horizontal correlation **only** (*Alexander* being the only outlier), and this clearly reflects that SVD with bias is including additional

factors to capture the biases, resulting in a larger number of factors. These additional factors are appearing as additional axes in the factorization visualization.

The most popular movies and best movies visualizations both show different behavior within the same model as well as between the two models.

Without bias (popular) – graph is more clustered and concentrated in the bottom right reflecting a stronger vertical gradient than horizontal correlation.

Without bias (best) – on the other hand, has a more left-skewed orientation with a strong vertical and horizontal presence and more widespread than the popular movies graph.

When comparing between both models, taking *Lord of the Rings* movies would setup an optimal example. With bias, the model places the 2003 sequel far away than the other two closely clustered sequels which could reflect a certain factors being considered that the without bias model is not taking into consideration. We conclude that as the without bias SVD model seems cluster the three sequels closer to each other and along a similar vertical quotient, and minimal horizontal differences. These two categories of visualizations again reinforce the idea that Surprise SVD with bias is working more efficiently (it theoretically should too!) than the without bias model by accounting for a larger set of factors to visualize these movies. Though, the differences are not entirely clear in the best movies case between the models, the with bias model still seems to evaluate way different vertical gradients for *Akeelah and the Bee* and *Louis CK* than without bias version. This cannot directly be correlated with an increased amount of latent features as both models' graphs seems to similarly widespread and not too tightly clustered.

Animated, Horror, Documentary Genres - with vs without bias

When examining the plots of the biased Surprise SVD for different movie genres, it is observed that the plots for horror and documentary movies are much more widespread, while the plots for animated movies are more tightly clustered. This suggests that the model uses different parameters for different genres. By incorporating the bias term in the model, it can account for user-specific and item-specific biases, such as popularity and rating. As a result, the model can focus more on the features of the movie rather than its popularity and accurately predict the user's interest. On the other hand, when looking at the plots for the unbiased Surprise SVD for different movie genres, it is observed that the points across all three genres are pretty concentrated. The points in the horror movies plot are almost split in half, suggesting that there could be features about each cluster of horror movies that the model is capturing. However, the model may still be influenced by the popularity and rating of the movie, resulting in a skew towards popular movies.

Overall, the biased Surprise SVD is expected to perform better than the unbiased Surprise SVD as it can remove the influence of popularity and rating and focus more on the features of the movie. However, both models still share a common trend where the more popular movies are more concentrated towards the right side of the x-axis, and the less popular movies are more spread throughout the graph. This still seems like a rough, loose analysis as the above observation is not constantly and religiously reflected by all graphs, which is not a bad thing as both axes are not intended to represent fixed values (overfitting is being prevented). However, we can also expect such a trend as popularity and audience reception greatly

affects the rating of a movie, and the biased Surprise SVD seems to account for this effect more accurately. Therefore, the biased Surprise SVD results in a better model that accurately represents the data.

Matrix Factorization Methods

Although we have a limited selection of methods and movie sets, we can extract general characteristics of each method applied to different movie sets. Since the actual results are difficult to analyze, we only consider the projection of the matrix factorization axes for two dimensions, however, one appropriate device for comparison is comparing clustering. Method 3 produces the best clustering (Figure 19 is a prime example) while Methods 1 and 2 produce fairly sparse plots. We also consider the horizontal and vertical alignment of movies with "Feature 1" and "Feature 2" axis. By inspection, it is clear that Method 1 and 2 are better at aligning movies with fixed values of Feature 2 (Figures 8 and 14). In Figure 8, we see that feature axis 1 and 2 are strongly orthogonal (in a PCA sense). we see this by observing that most movies align on $F_1 = 1.5$ and $F_1 = 0.75$. These are qualities that are strongly present in our interpretation of the data.

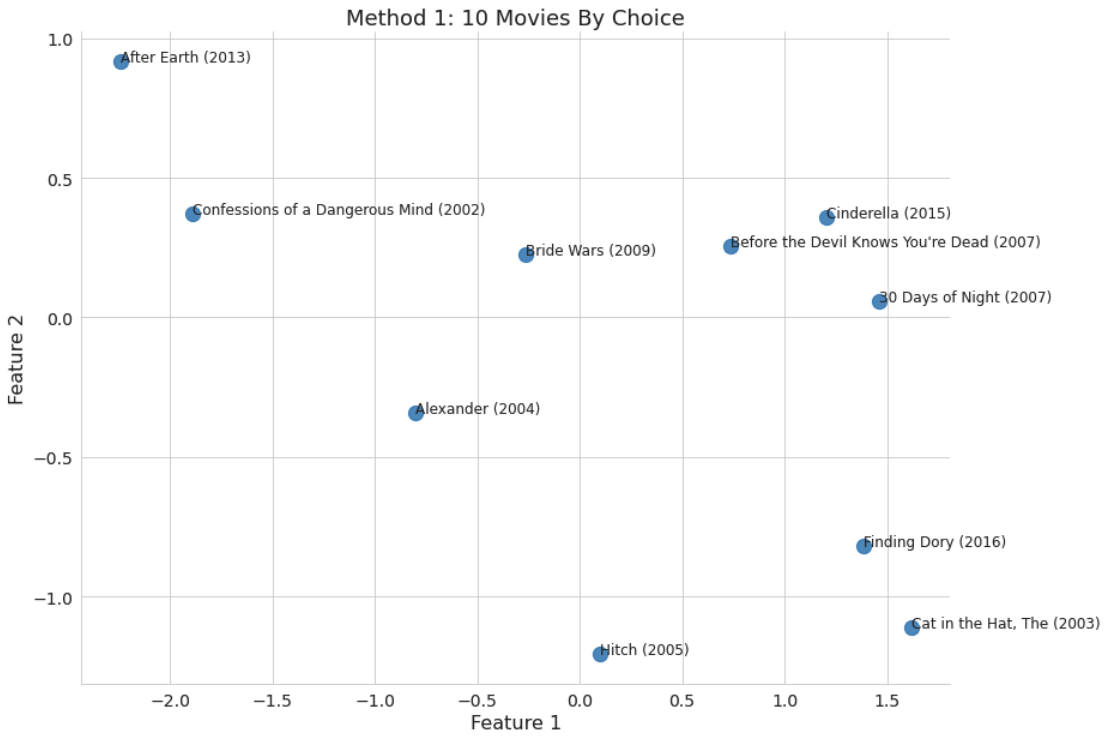

Figure 7: **HW5 A:** Any ten movies of your choice from the MovieLens dataset.

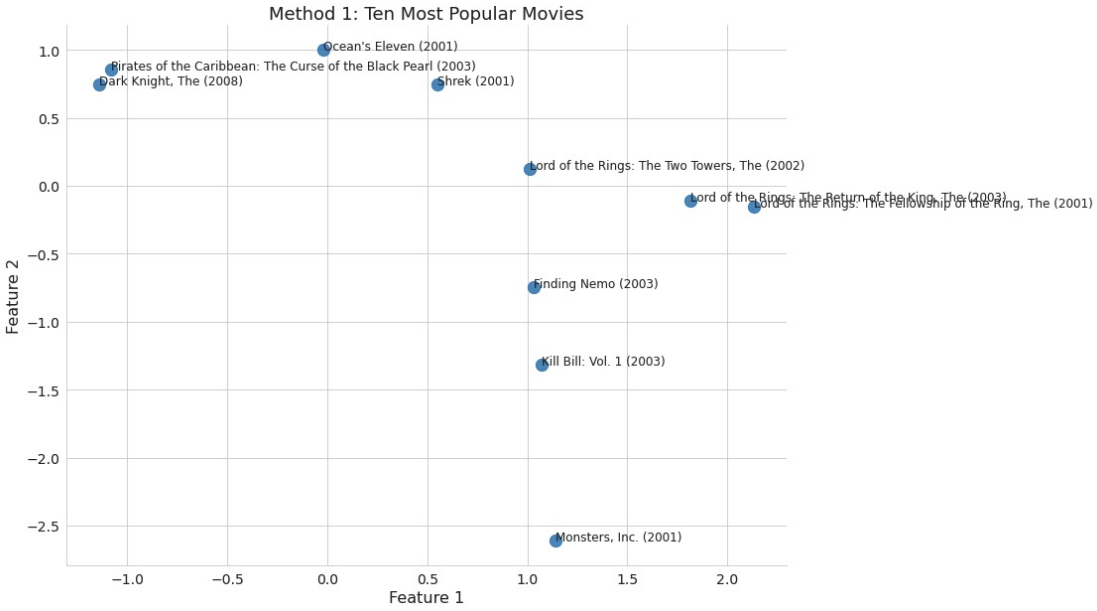

Figure 8: **HW5 B:** The ten most popular movies (movies which have received the most ratings).

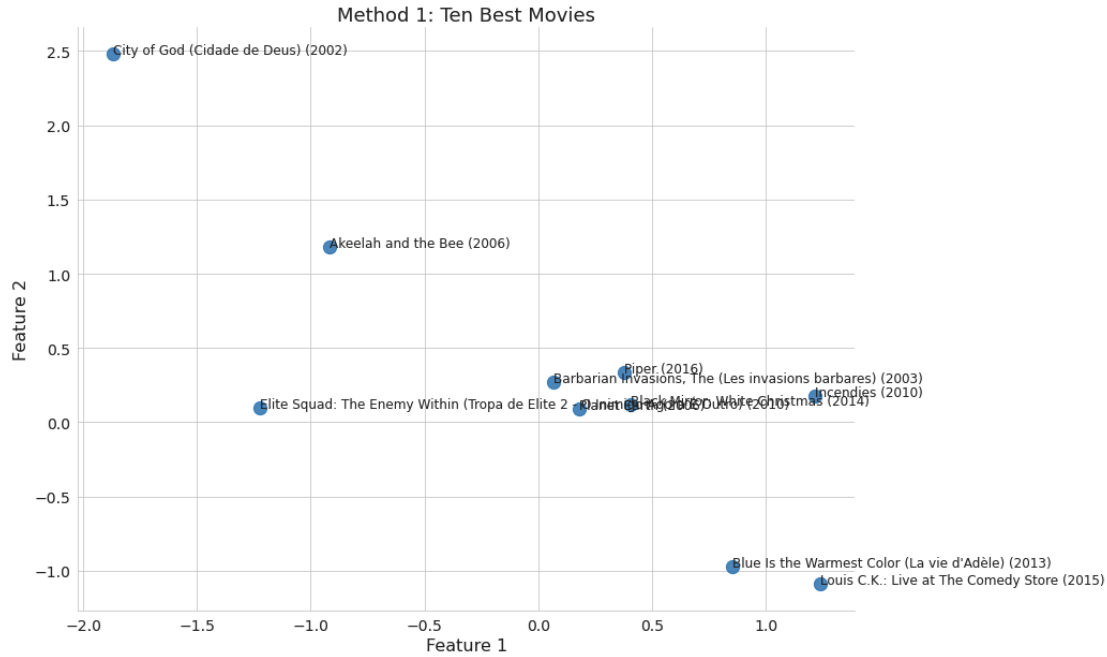

Figure 9: **HW5 C:** The ten best movies (movies with the highest average ratings).

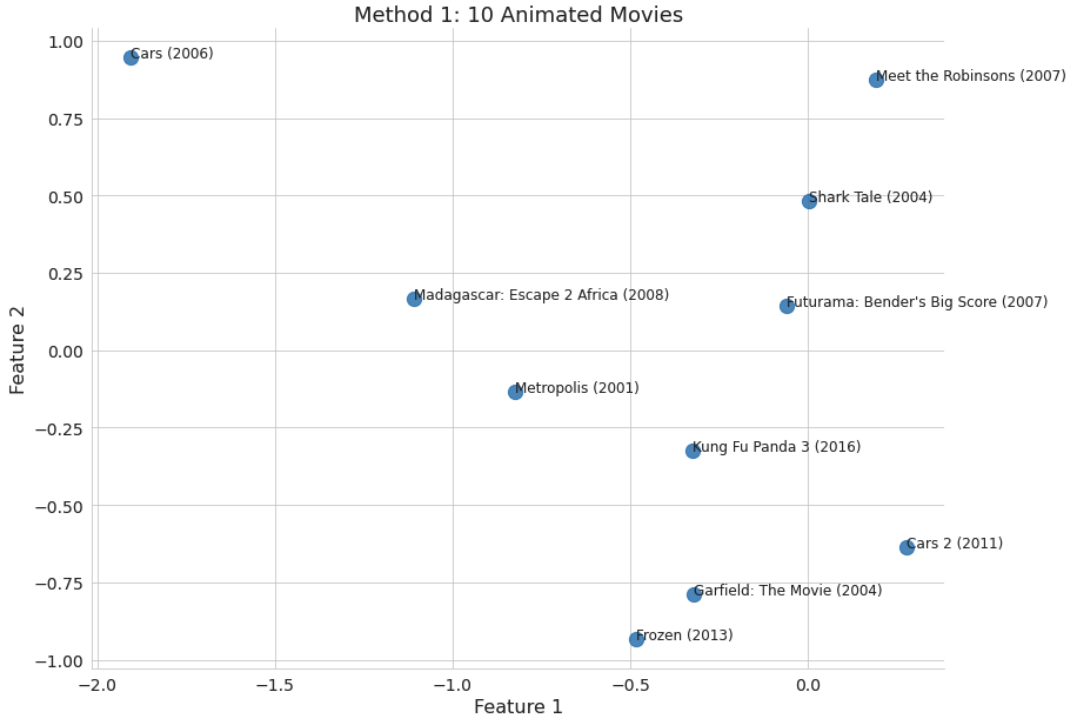

Figure 10: **HW5 D1:** Ten movies from the genre Animation.

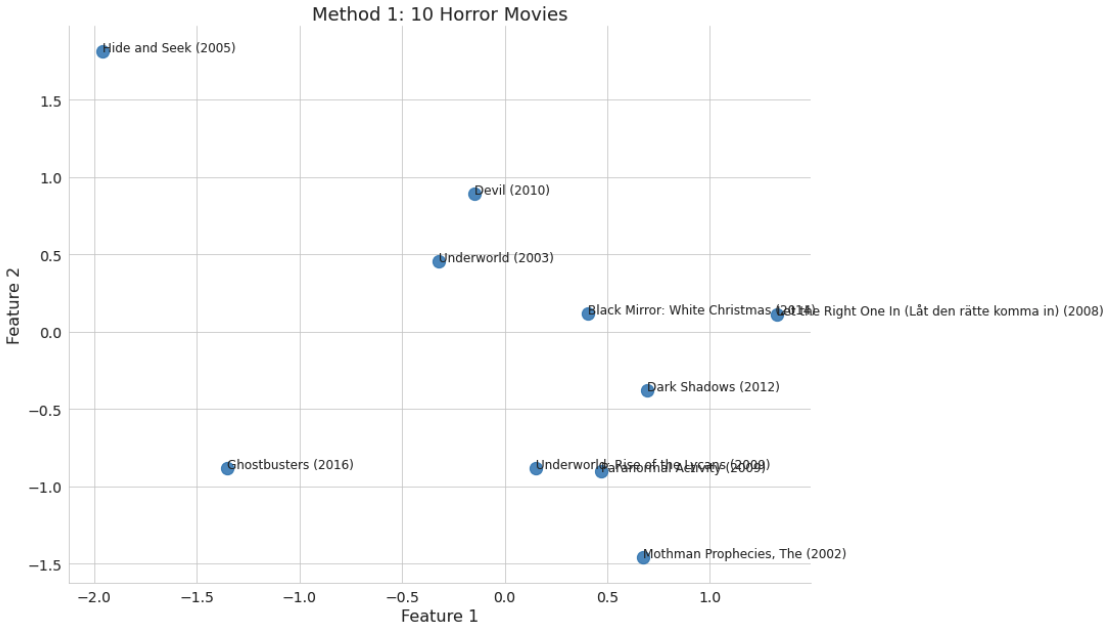

Figure 11: **HW5 D2:** Ten movies from the genre Horror.

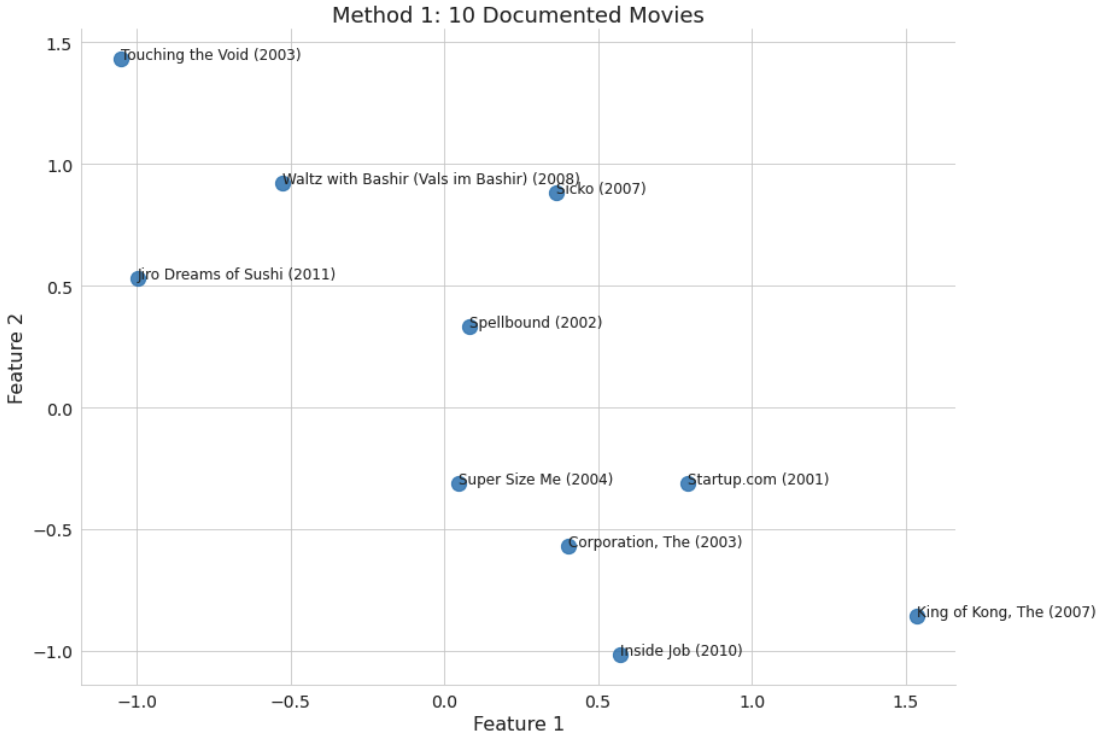

Figure 12: **HW5 D3:** Ten movies from the genre Documentary.

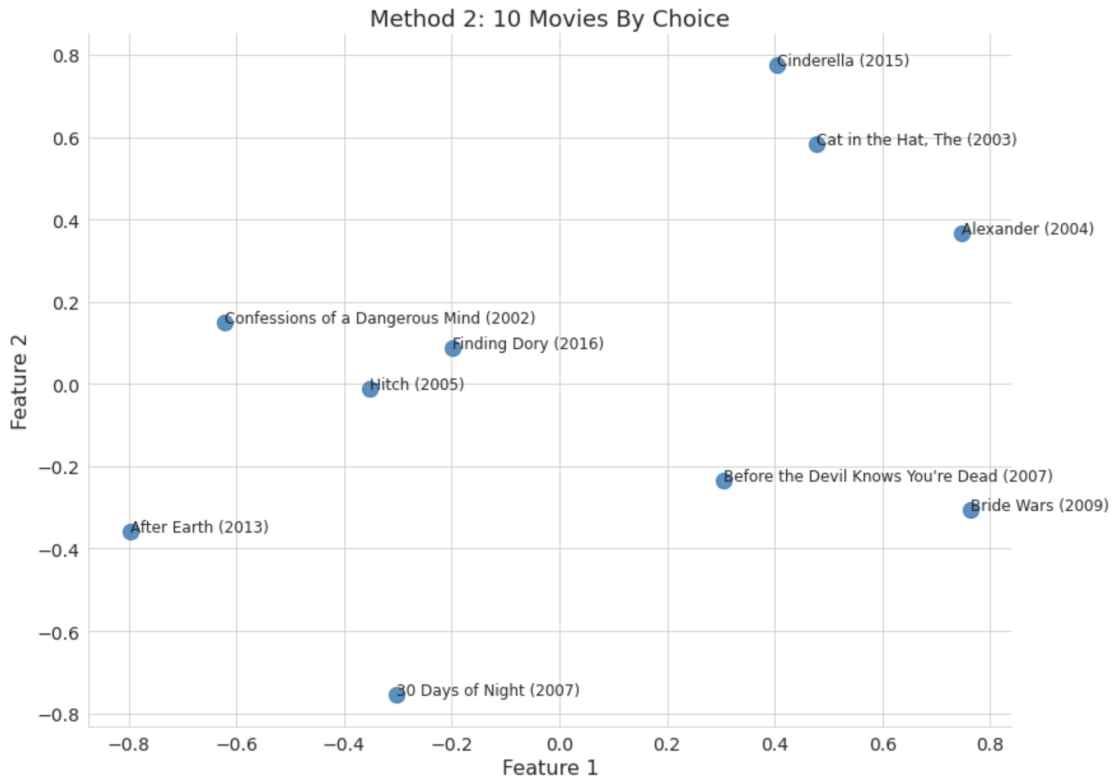

Figure 13: **Surprise SVD with Bias A:** Any ten movies of your choice from the MovieLens dataset.

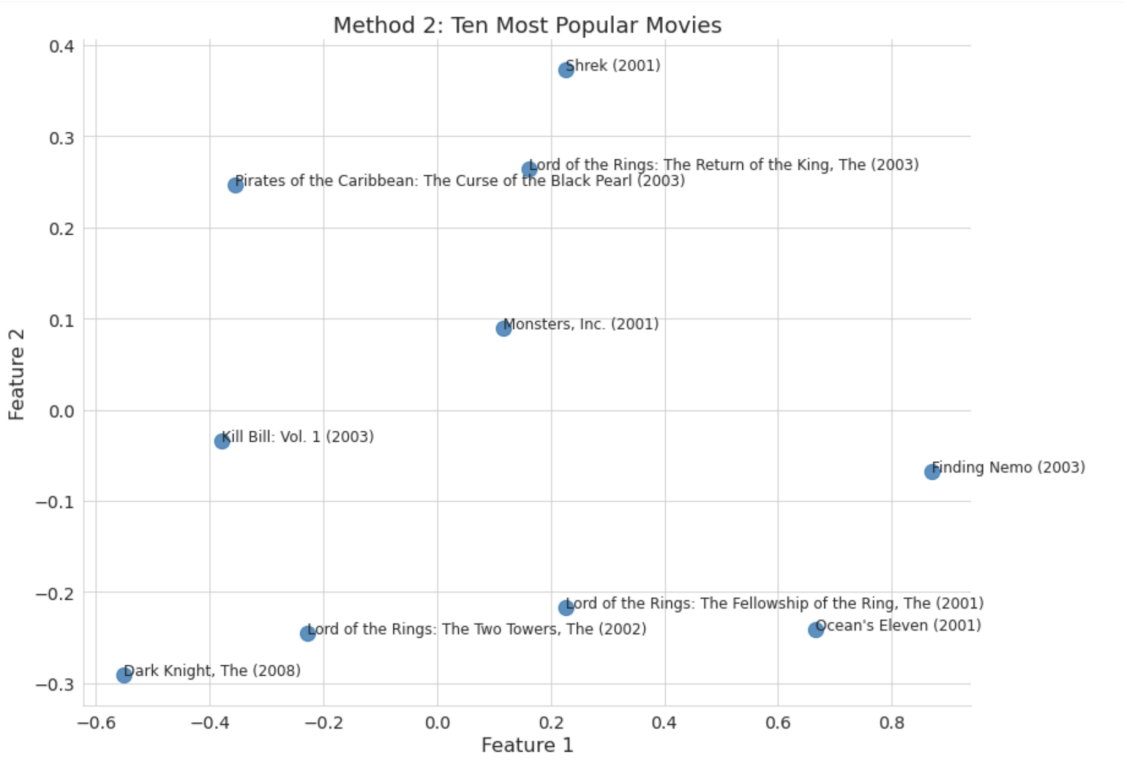

Figure 14: **Surprise SVD with Bias B:** The ten most popular movies (movies which have received the most ratings).

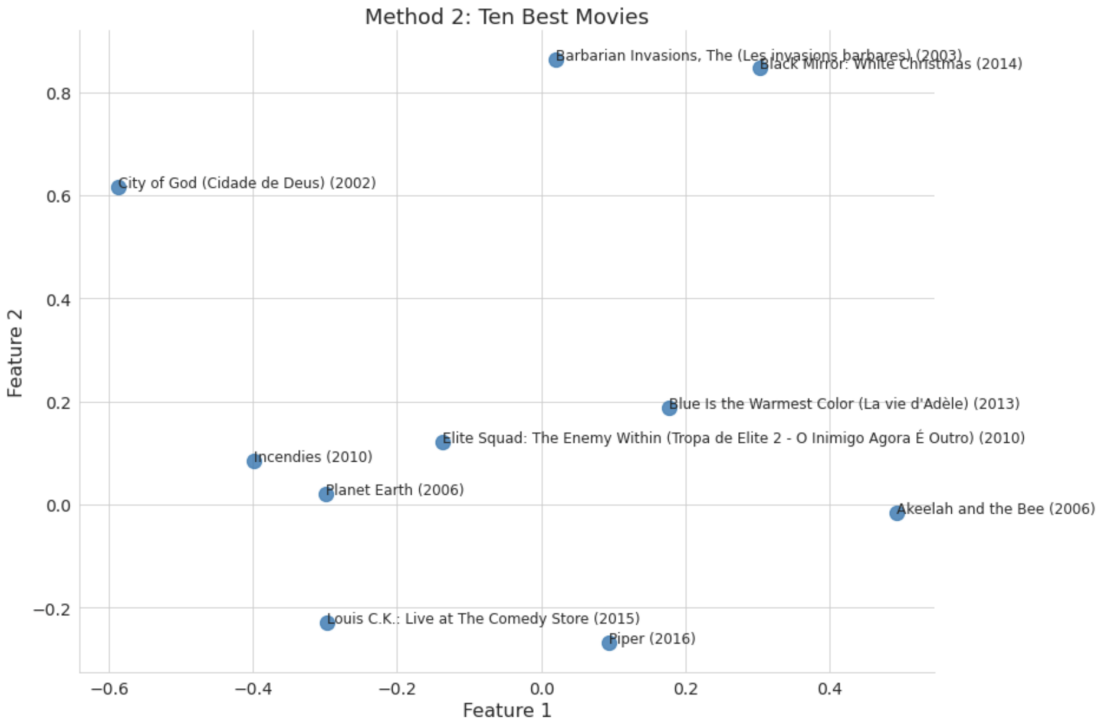

Figure 15: **Surprise SVD with Bias C:** The ten best movies (movies with the highest average ratings).

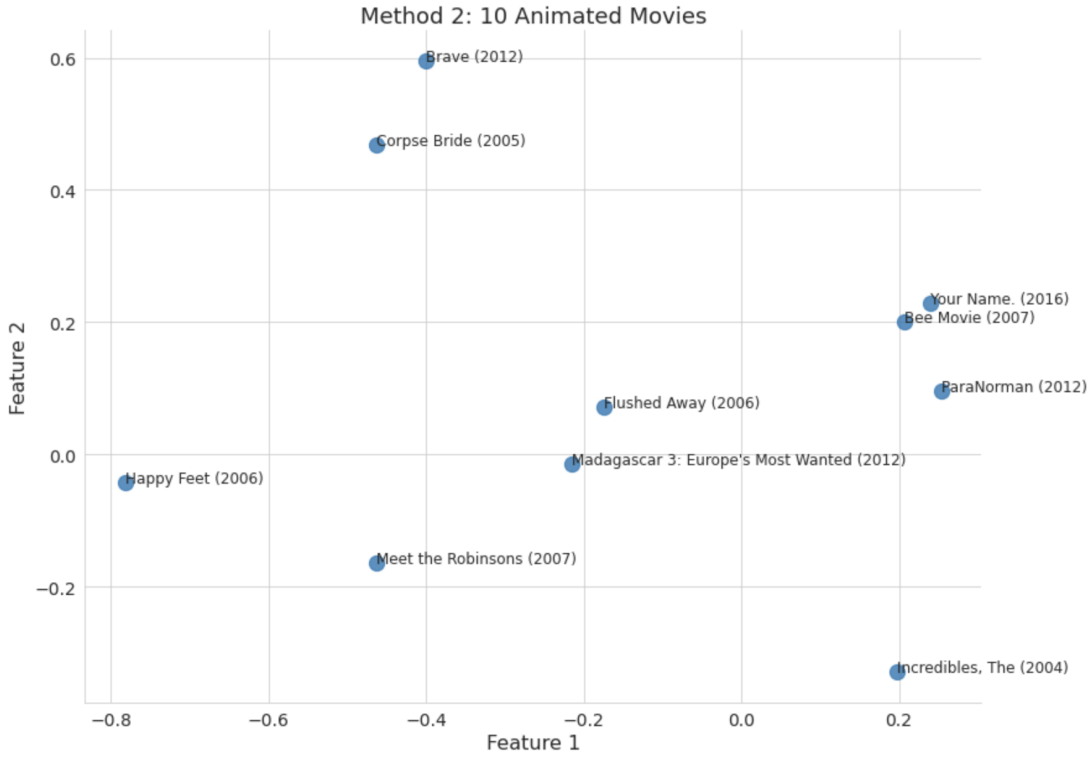

Figure 16: **Surprise SVD with Bias D1:** Ten movies from the genre Animation.

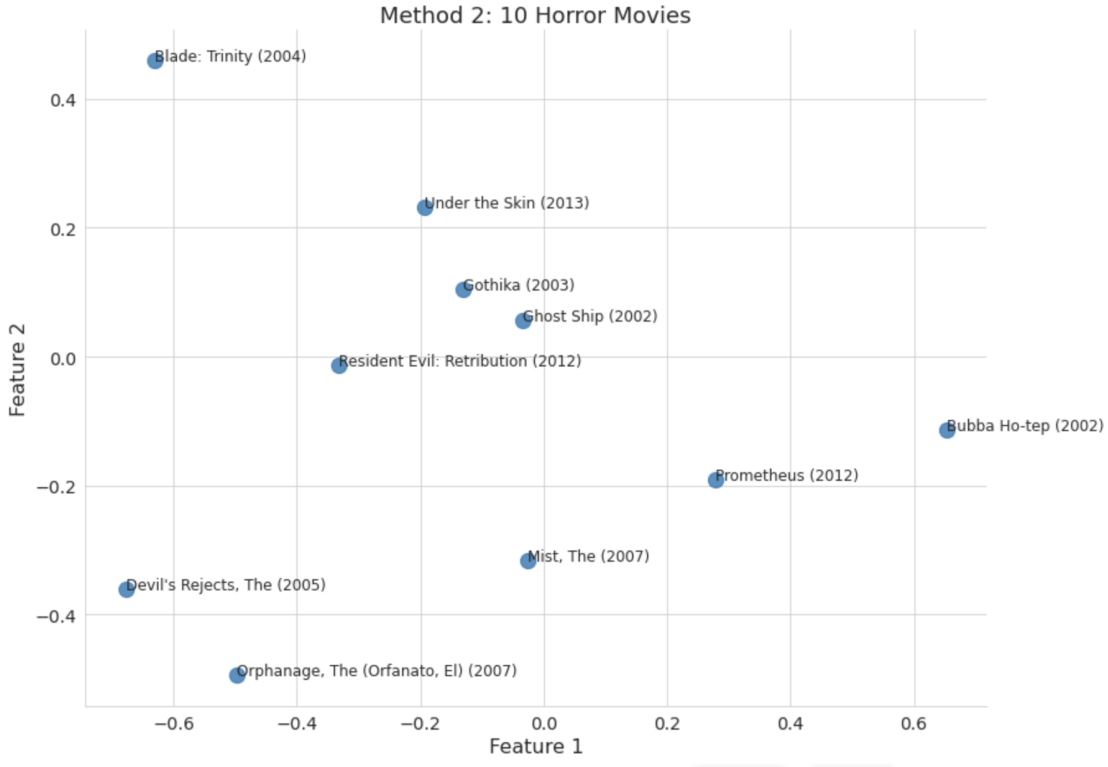

Figure 17: **Surprise SVD with Bias D2:** Ten movies from the genre Horror.

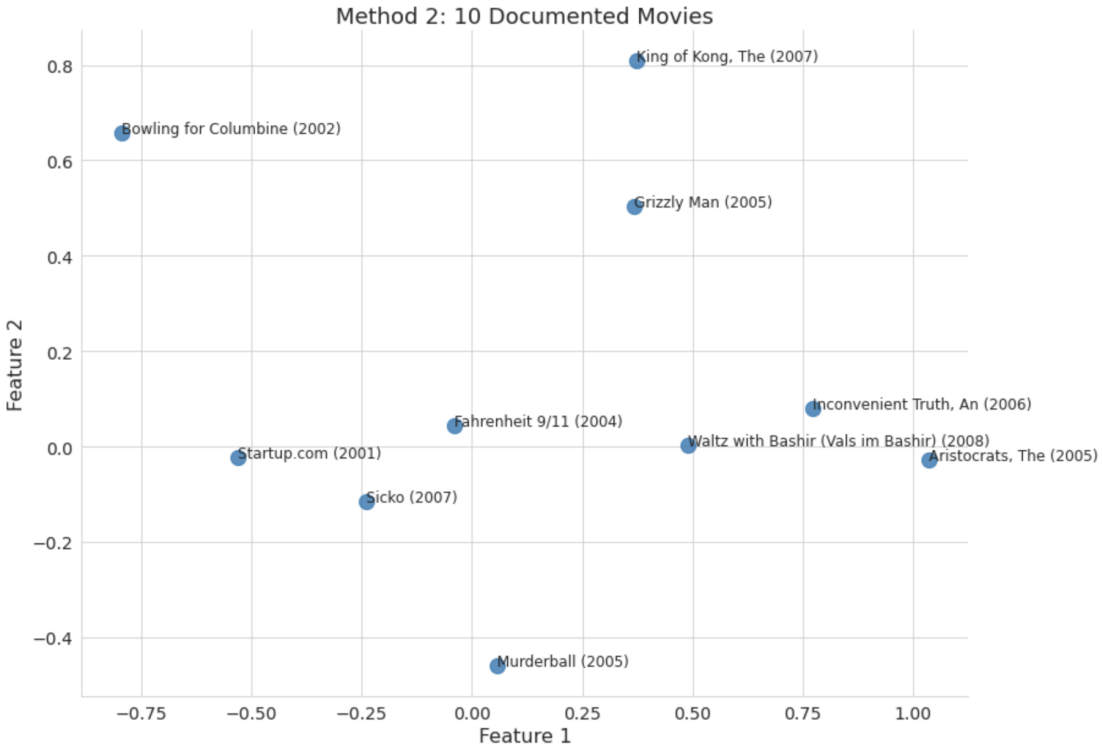

Figure 18: **Surprise SVD with Bias D3:** Ten movies from the genre Documentary.

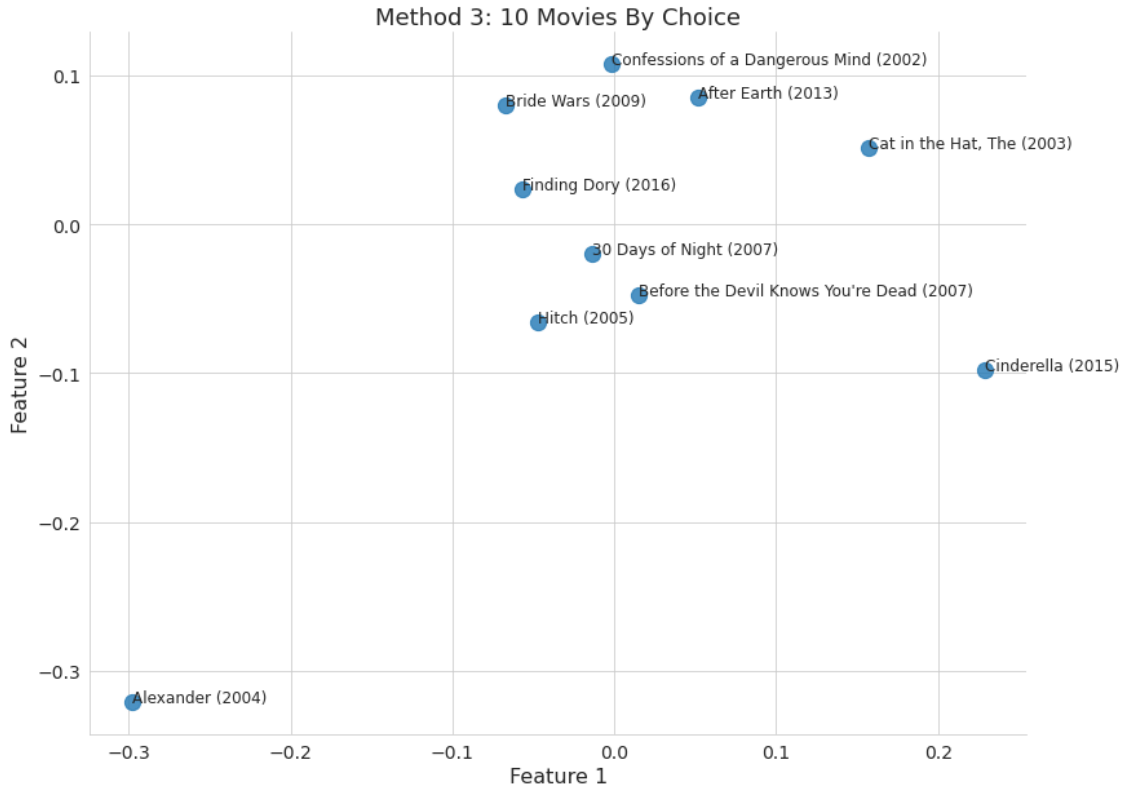

Figure 19: **Suprise SVD no Bias A:** Any ten movies of your choice from the MovieLens dataset.

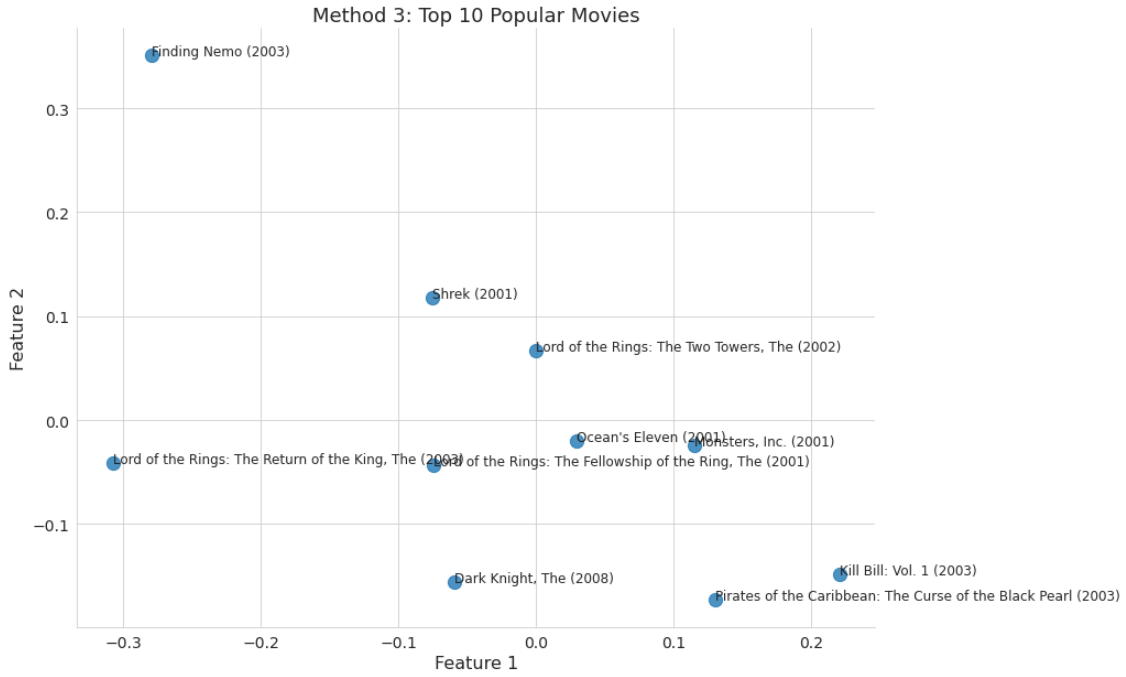

Figure 20: **Suprise SVD no Bias B:** The ten most popular movies (movies which have received the most ratings).

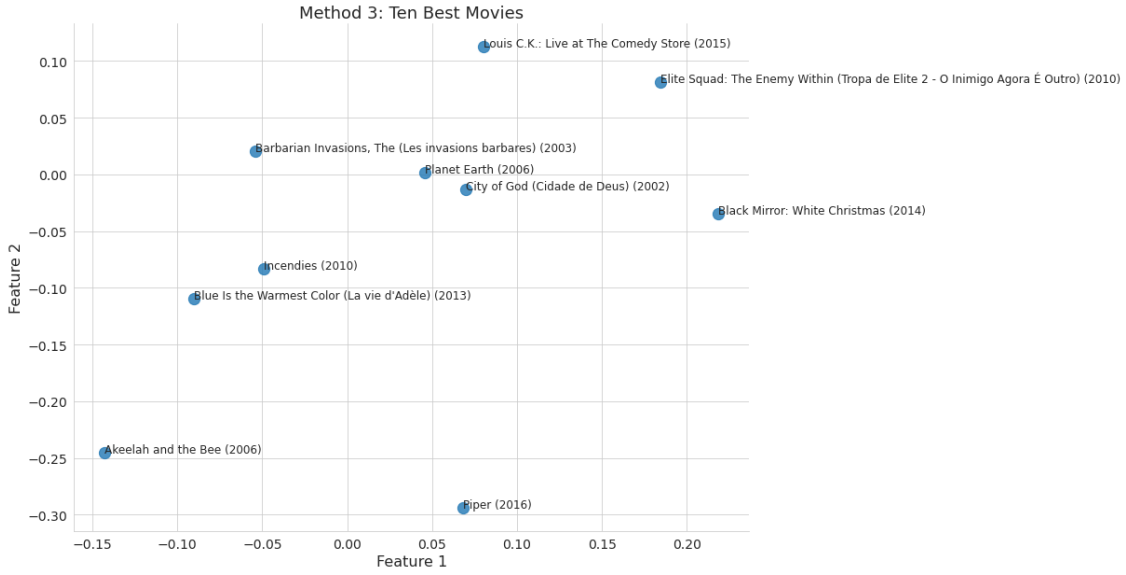

Figure 21: **Suprise SVD no Bias C:** The ten best movies (movies with the highest average ratings).

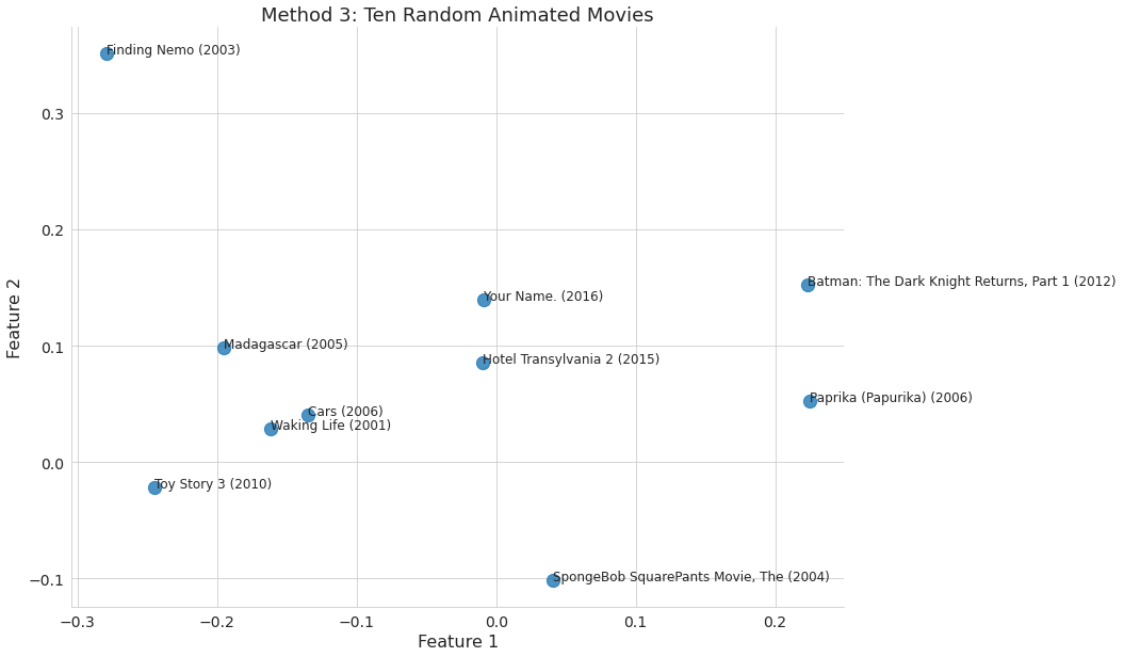

Figure 22: **Suprise SVD no Bias D1:** Ten movies from the genre Animation.

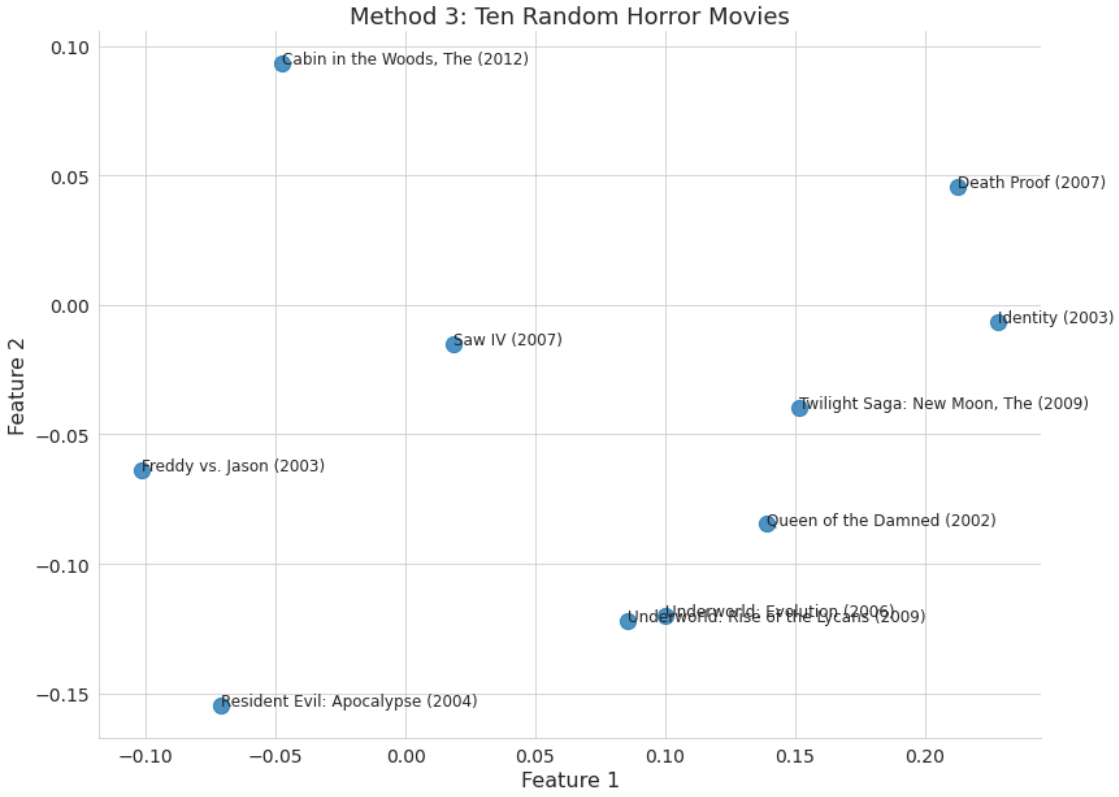

Figure 23: **Suprise SVD no Bias D2:** Ten movies from the genre Horror.

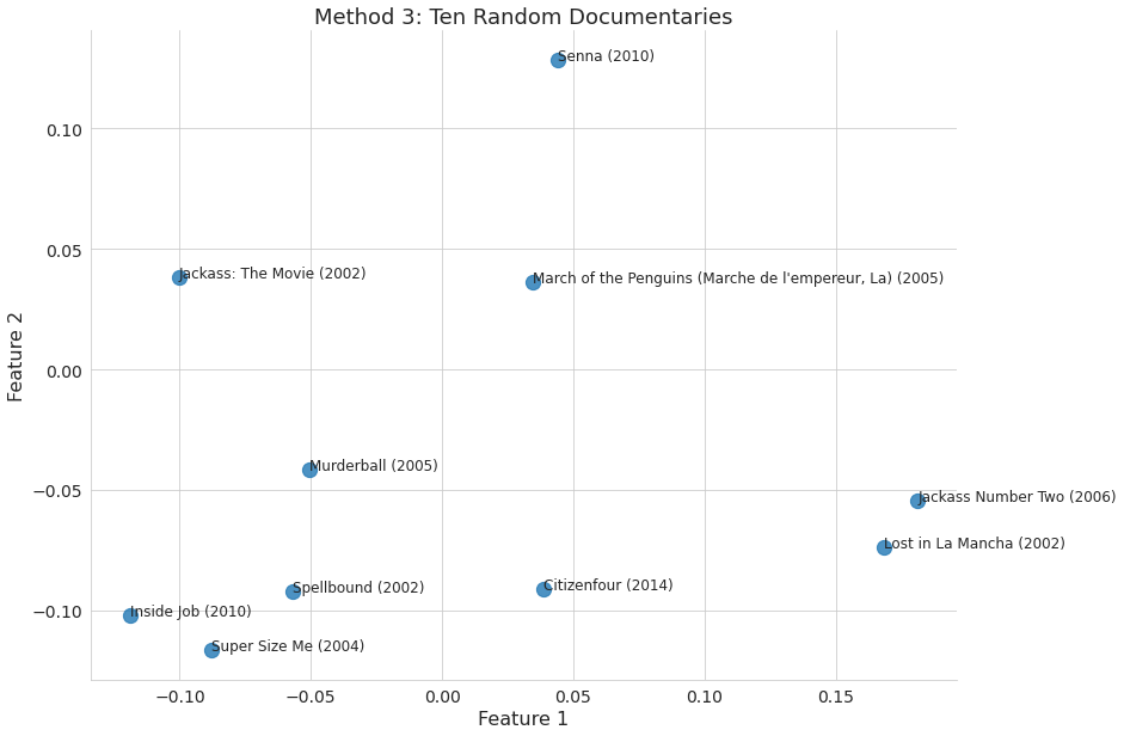

Figure 24: **Suprise SVD no Bias D3:** Ten movies from the genre Documentary.