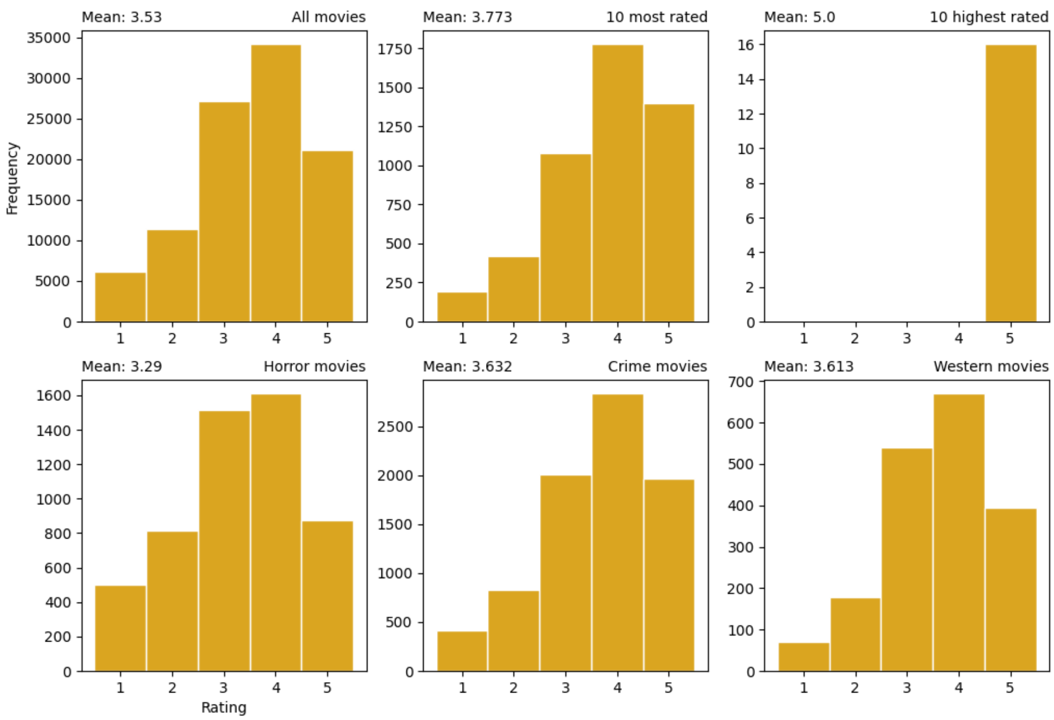
1 Introduction

- Team Name: jc
- Group members: Chenning Xu, Joahan Castañeda Jaimes
- Division of labor:
- Packages used: suprise, numpy, matplotlib

2 Basic Visualizations

Our group produced the following histograms for the MovieLens dataset:



We noticed that the most common rating that was given to a film was a rating of 4. The distribution as a whole is left skewed with a mean value of 3.53. One might have suspected that a rating scale of 1 to 5 would yield a mean rating of 3, but it appears that this rating distribution does not obey a normal distribution.

The shape of the ratings distribution for the most popular movies is similar to the rating distribution of the whole dataset. The mean is slightly higher and there are now more 5 ratings than 3 ratings. Therefore the popular movies tend to have higher ratings as a whole, but not by an absurdly noticable amount. This is what our group expected to see.

There are only ratings of 5 on this histogram for the 10 highest rated movies since the movies with the highest ratings are those which have had very few users rate them and rate them very highly (as is indicated by the lower frequency counts). This is what we would expect.

For our three genres, we selected “Crime,” “Horror,” and “Western.” Crime movies were the most popular and were marginally the highest rated genre of the three that we chose. Horror moves were the second most popular and were by far the lowest rated genre of the three. Finally, Western moves were the least popular but were comparable in mean rating to the movies in the Crime genre.

3 Matrix Factorization Visualizations

Model training and explanation

In the first model that we trained, we following the factorization procedure that was outlined in Set 5. To derive the coefficients of the matricies $U \in \mathbb{R}^{M \times K}$ and $V \in \mathbb{R}^{N \times K}$, we minimize the regularized square error as defined by the following objective:

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

where u_i^T and v_j^T are the row components of U and V and the terms inside the first parentheses correspond to the Frobenium norms of the two matricies. This gives rise to a final matrix $Y = UV^T$ where $y_{ij} \approx u_i^T v_j$. To minimize this objective, stochastic gradient descent is utilizied with the following rule updates:

$$u_i = u_i - \eta \partial_{u_i}, \tag{2}$$
$$v_j = v_j - \eta \partial_{v_j}. \tag{3}$$

Stochastic gradient descent is stopped once the relative loss reduction compared to the first epoch is beneath some threshold ϵ:

$$\Delta_{t-1,t}/\Delta_{0,1} \leq \epsilon. \tag{4}$$

In the second model that we trained, we introduced bias terms a and b to the procedure outlined in Set 5. With these new terms, we now minimize the following objective:

$$\arg\min{}_{U,V} \frac{\lambda}{2}\left(||U||_F^2 + ||V||_F^2\right) + \frac{1}{2}\sum_{i,j}(y_{ij} - (u_i^T v_j + a_i + b_j))^2. \tag{5}$$

We otherwise use the same gradient descent procedure as is used in the first model, only now we must include the update rules for a and b:

$$a_i = a_i - \eta \partial_{a_i}, \tag{6}$$
$$b_j = b_j - \eta \partial_{b_j}. \tag{7}$$

For the off-the-shelf implementation, we used the `surprise` package.

Model parameters and performance

Our model performances are listed in the following table:

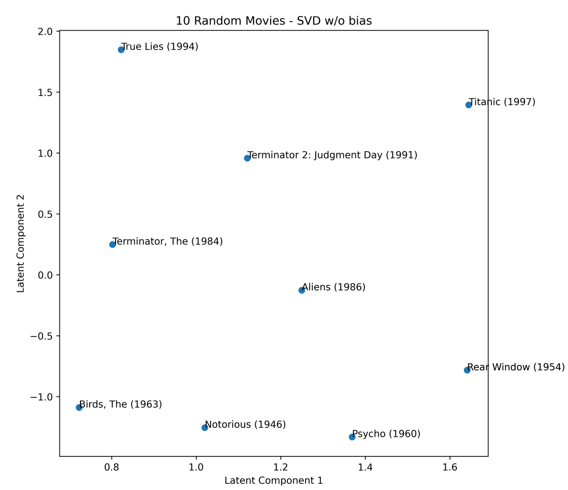
Model	E_{in}	E_{out}
SVD w/o bias	0.336	0.454
SVD w/ bias	0.276	0.446
SVD (`surprise`)	0.873	0.920

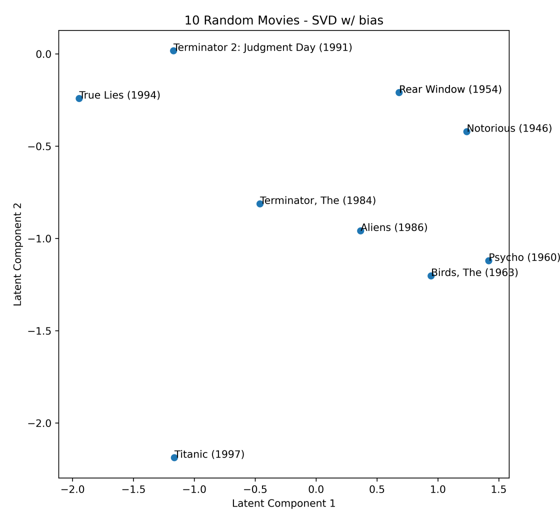
For the standard SVD model with no bias terms, we tuned our model to an early stopping criterion of $\epsilon = 0.001$, a learning rate of 0.03, and a regularization value of 0.1. These values were used as they minimized overfitting, as was evident in Problem 2 of Homework Set 5. We found that very similar parameters were effective to the SVD with the nearly introduced bias terms. We used the same parameters, with the exception of the regularization parameter which we found to be most effective at 0.09.

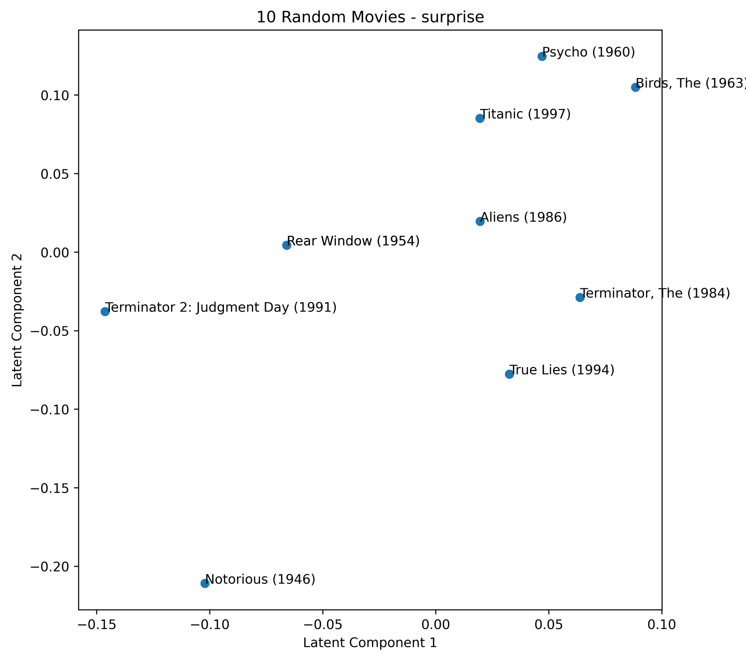
Finally for the `surprise` implementation of SVD, we performed a GridSearch of the model parameter space to tune our model on the dataset. We used default values for the `surprise` implementation of SVD except for the learning rate and regularization parameter, which we found to be best selected as 0.0264 and 0.113, respectively. This implementation has the worst error for both the training set and testing set.

4 Model visualizations

Random ten movies



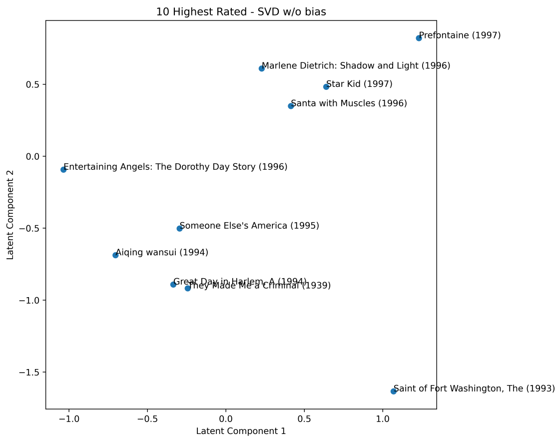


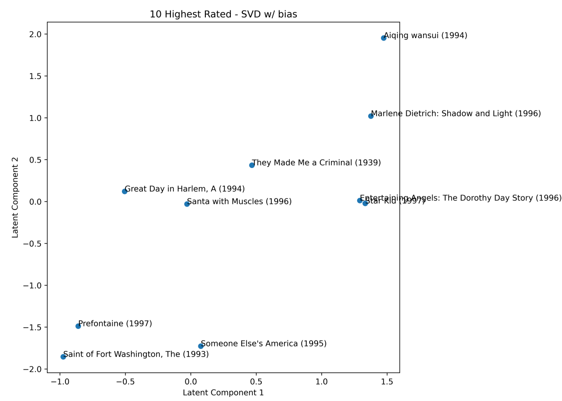


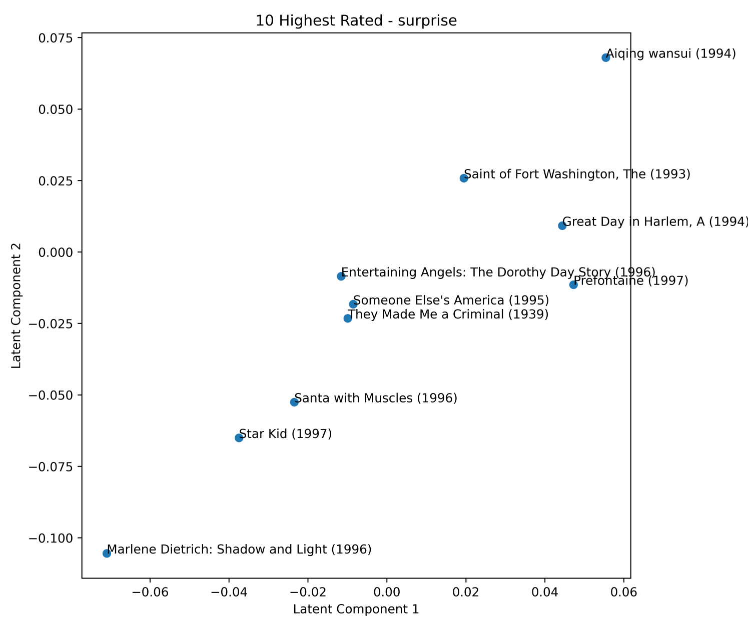
These are the films that we used for our Piazza post. 10 films were selected, 6 of which were Alfred Hitchcock filsm and 4 of which are James Cameron films. *Aliens*, *True Lies*, *The Terminator*, and *The Terminator 2* are roughly clustered together in the SVD plots of our own implementation with the association being marginally stronger in the model with the bias terms. Our SVD with bias identified *Titanic* as an outlier, something that may have to do with its extreme popularity. This model also captures the similar themes between *The Birds* and *Psycho*, both being Mystery/Horror films, something that the `surprise` model also appears to capture.

In the `surprise` model, `Notorious` is captured as an outlier, perhaps owing to its abnormally old age among the rest of the films. Otherwise, we don't note a significant difference between it and the other SVD implementations. Perhaps we could argue that the SVD plot without the bias terms has spread out the points more than the other models, but that's about it.

Ten most highly rated movies

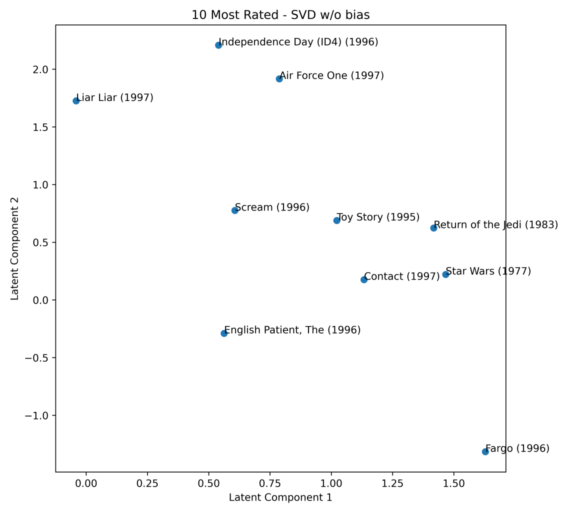


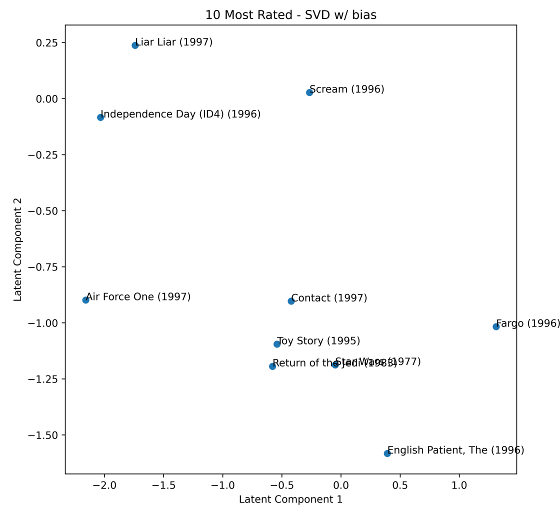


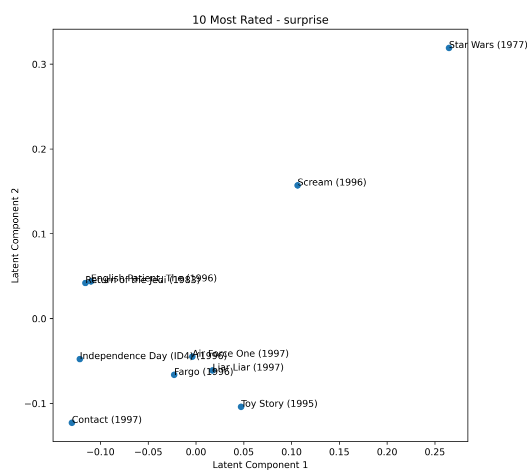


These films do not have a significant amount of ratings and may struggle with correlating well with other similar films in these plots. Despite this, *aiqing wansui* stands out as an outlier in the biased SVD implementation and the `surprise` implementation. This film is Taiwanese, the only foreign film in this set of films so that could be why its placed in this particular way. We don't really note any significant correlations between films in any implementation, understandably so since these films don't really have as many data points to train on.

Ten most popular movies



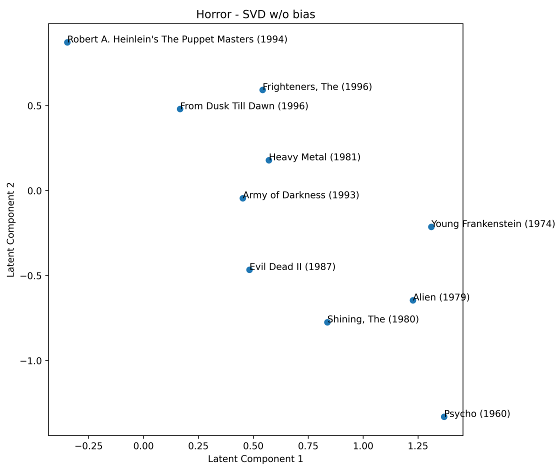


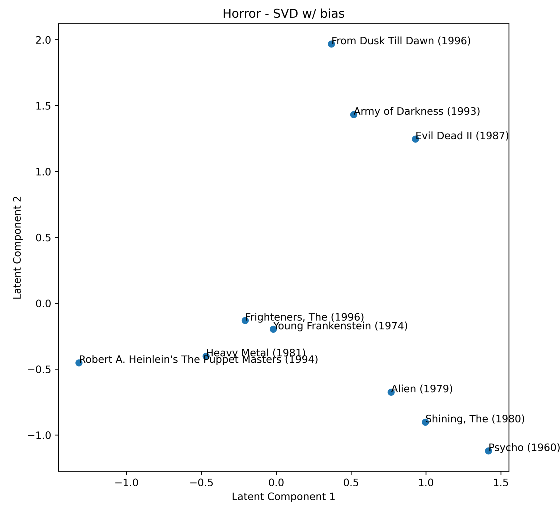


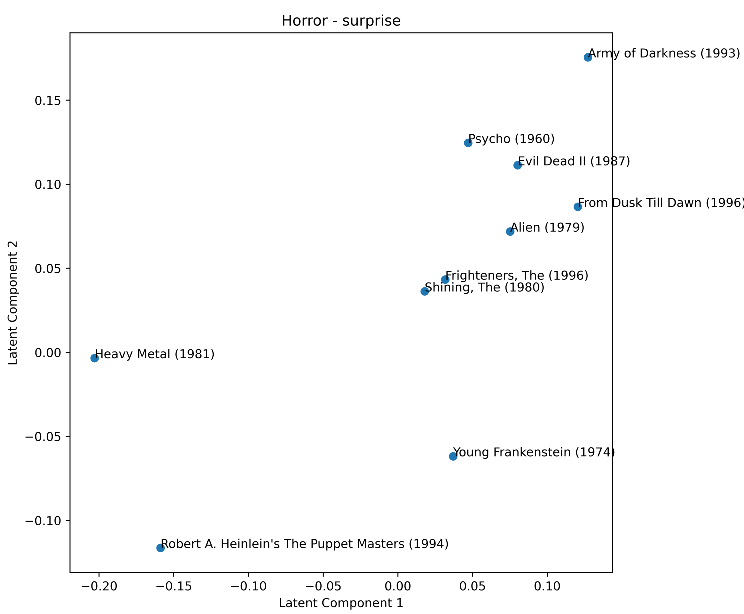
The `surprise` implementatin of SVD does not appear to make as much use of the parameter space as either of the other two SVD implementations that we made. The placement of the different films does not make the most sense and it surely reflected of the subpar performance that `surprise` yielded on the test set.

When looking at the our own SVD implementations, most movies have been shifted around by a decent amount but largely remain in the same quadrants as they had been in previously. *Air Force One* and *The English Patient* had the greatest shifts with *Air Force One* moving to an entirely new quadrant, but otherwise most points remain roughly in place. Some of the positions of these points make sense, for instance the two Star Wars films being close to each other in both plots. Meanwhile other point positions seem more questionable, such as *Independence Day* and *Contact* not being near each other despite both being sci-fi films.

Horror movies



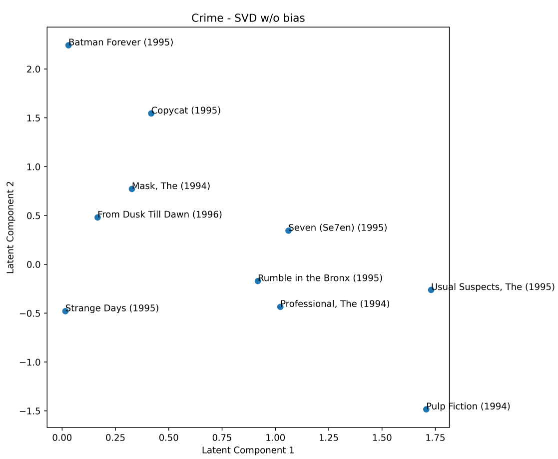


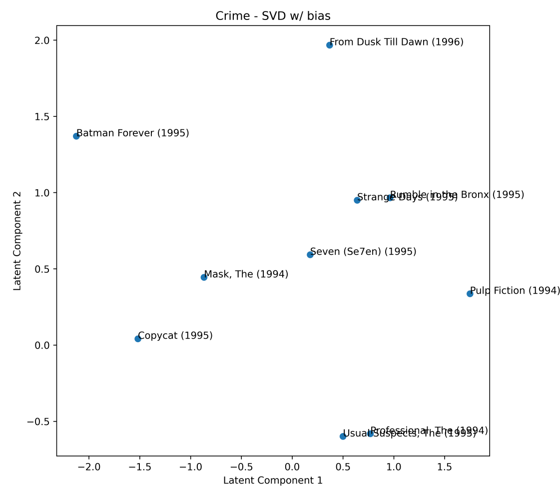


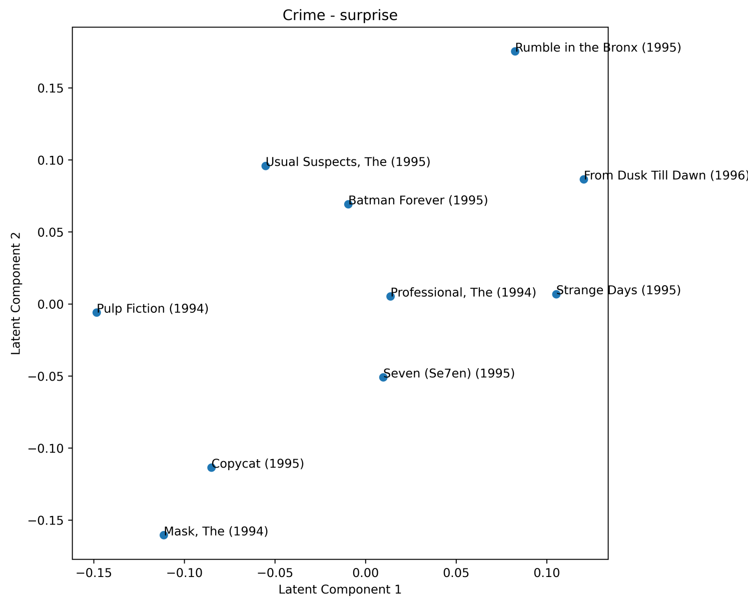
The horror films appear to be subject to more clustering than the other film groups we've looked at. The `surprise` implementation is most guilty of this and does not appear to make a clear distinction between different subgenres of horror. The top right quadrant of our plot puts together horror films with comedic, mystery, and action elements together.

The other two SVD implementations seem to do a better job at clustering cimilar movies. For instance, we would expect *The Shining* and *Psycho* to be near each other on the basis that they are both horror/mystery thrillers. *Alien*, also being in the horror/mystery category is located near these other two movies. *Army of Darkness* and *Evil Dead II* are also near each other in both plots as horror/comedy films. The remaining films don't seem to be scattered too coherently. Interestingly, the SVD model without the bias terms displays these points as a diagonal continuum across the parameter space while the SVD model with the bias terms is more willing to cluster movies in an isolated fashion.

Crime movies

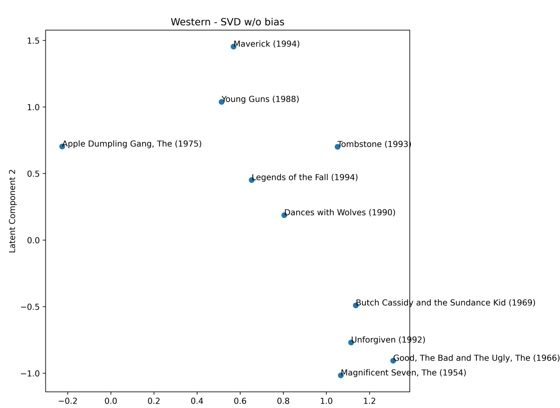


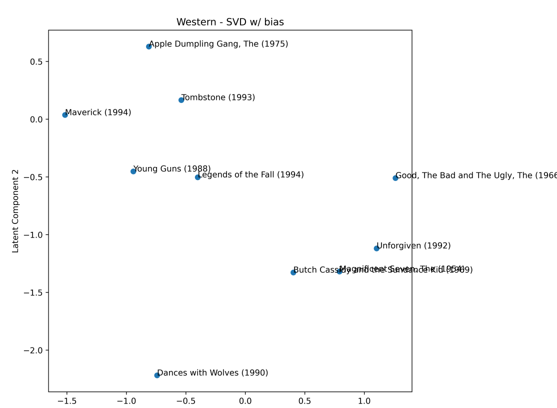




All of our models seemed to struggle with the crime movies than any other the other sets of films, though some trends are common among them. *Pulp Fiction* manages to be somewhat of an outlier in every plot, maybe due to its relative popularity compared to the other films that are being included. *The Profession* and *The Usual Suspects* are close to each other in every plot, though none of the other really seem to have similar clustering tendencies. There are three comedy/crime films: *Batman Forever*, *The Mask*, and *Rumble in the Bronx*, none of which seem to group up in any plot.

Western movies





Many Western films are also crime-fiction films which could be why films such as *Unforgiven*, *Butch Cassidy and the Sundance Kid*, *The Good, the Bad, and the Ugly*, and *The Magnificent Seven*, are all close to each other in the two SVD plots corresponding to our implementation. *The Apple Dumpling Gang* and *Maverick* are also near each other in both of these plots which is also reasonable considering that both films have comedy elements.

The *surprise* implementation makes it's more convincing predictions in this set of films. The films that are clustered together in our implementation are also captured by *surprise*.

5 Piazza post

https://piazza.com/class/m4xf8vvl4206zc/post/220

6 LLM Usage