Contents

Abstract

This monograph describes author's large experimental work on one machine learning task – prediction of movie ratings in the Netflix Prize dataset. The main objective of the experiments was to obtain maximally accurate prediction, as evaluated by hold-out RMSE, but also important was the perspective of applying the developed methods in recommender systems. The publication has two goals: summarizing the understanding of the subject due to the published work of many people on the same task, and presenting some novel insights. Reaching a good understanding of one task and one dataset gives hope to generalize on other prediction tasks, as similar challenges recur in analyses of any datasets.

The idea of collaborative filtering is to make use of relations between tasks (users in our data), and between task attributes (items in our data). Collaborative filtering methods are used in recommender systems to calculate personalized recommendations, or in other words, to identify items preferred by a particular user. To realize that goal, a good intermediate task is prediction of user ratings, and the most accurate models for this task are based on dimensionality reduction, describing each item by a small number of variables, which can be seen as automatically learned analogues of movie genres, and a small number of variables describes each user's taste. One the most accurate models, regularized SVD, was analyzed more closely, and the assumptions of that model, such as the single-variable output, combining hidden variables by multiplication, and using Gaussian priors, were critically examined. In addition, an interpretation of the learned features by naming new movie genres has been proposed.

To learn the parameters in the developed models the best predictive accuracy was obtained by using different degrees of approximation of the Bayesian approach, from MCMC and Variational Bayes, to neural-networks-like simplifications. When identifying the model, that is, while approaching the unknown probabilistic model that generated the data, good engineering practice was maintaining a blend of an ensemble of many accurate, but varied methods. Blends of large ensembles also gave the best reached accuracy, indicating that, despite the large combined effort of many people, the process of model identification for the analyzed data remained largely unfinished, which is probably an unavoidable situation in an analysis of real-life datasets.

The work is complemented by giving heuristics adapting rating prediction to generate lists of recommendations, heuristics for cold-start situations, and descriptions of two SVD-based recommender systems.

I described the methods realizing the basic probabilistic matrix factorization model. They can be further improved by using time effects (section 4.3.6), better priors on user preferences (section 4.3.5), and using more refined global effects than the simple biases (section 4.3.1).

4.3.4 Examination of learned parameters

Matrix factorization methods perform dimensionality reduction and represent movies in the form of a short vector of learned parameters, which can be understood as automatically learned hidden genre representation. One might wonder, what is the relationship between the automatically learned genres and the genres distinguished by humans, such as action movie, comedy, horror, etc. The learned features can represent unnamed genres or may be a rotation, or other linear combination, of features intuitively understood by humans. For example, if movies were represented in two dimensions, say, on a plane spanned by two genres: "action" on X axis and "comedy" on Y axis (without defining those genres precisely), the basis vectors can be rotated and the movies can be represented in coordinates (1*action - 1*comedy, 1*action + 1*comedy), forming an equivalent matrix factorization after analogously rotating the user features. Different learning algorithms for sparse, regularized matrix factorization can result in different rotations of the solution, leading also to small differences in predictive accuracy. If features are learned sequentially, with each feature fully learned before moving on to learning the next ones, in the resulting factorization features will be ordered from the largest magnitudes, similarly as in the standard SVD algorithm from linear algebra, where features (singular vectors) are sorted according to their singular values.

I attempted to interpret a few features with the largest magnitudes as new movie genres. Features come from a VB version of regularized SVD (see section 4.3.3) with 32 features, fully learning one feature at a time, and with learning all features repeated three times. The experiments were backed by examining the standard genre classification and lists of keywords, both coming from the IMDb database correlated with the Netflix titles. It turned out, that the most meaningful SVD features encode concepts such as personality types, value systems, emotions, etc.

Figures 30 and 31 show the learned 32-element vectors of parameters $\tilde{\mathbf{v}}_\mathbf{j}$ for 35 example movies and TV series, chosen among the 4000 most frequently rated in the Netflix dataset. The vectors $\tilde{\mathbf{v}}_\mathbf{j}$ are outputs of a VB SVD, normalized to length one (normalized features were used in all experiments in this section). Next to every movie on the left side, two movies are shown on the right side for comparison: one positively correlated, and one negatively correlated (in a sense of the dot product of the normalized vectors $\tilde{\mathbf{v}}_\mathbf{j}^T \tilde{\mathbf{v}}_{\mathbf{j_2}}$).

Table 17 shows maximal and minimal values of the first six features among the 35 movies listed earlier in figures 30 and 31. Because of the chosen sequential way of learning the SVD model, the first features explain most of the variability in data. They correspond more or less to the largest singular values in a linear algebra SVD (but not precisely, because there is no orthogonality property, due to the presence of missing data and the use of regularization). The column "Rank" in the table 17 denotes the frequency rank – the ordering when the movies are sorted by support (by the number of ratings).

Figure 30: Example movies with their learned feature values, part 1

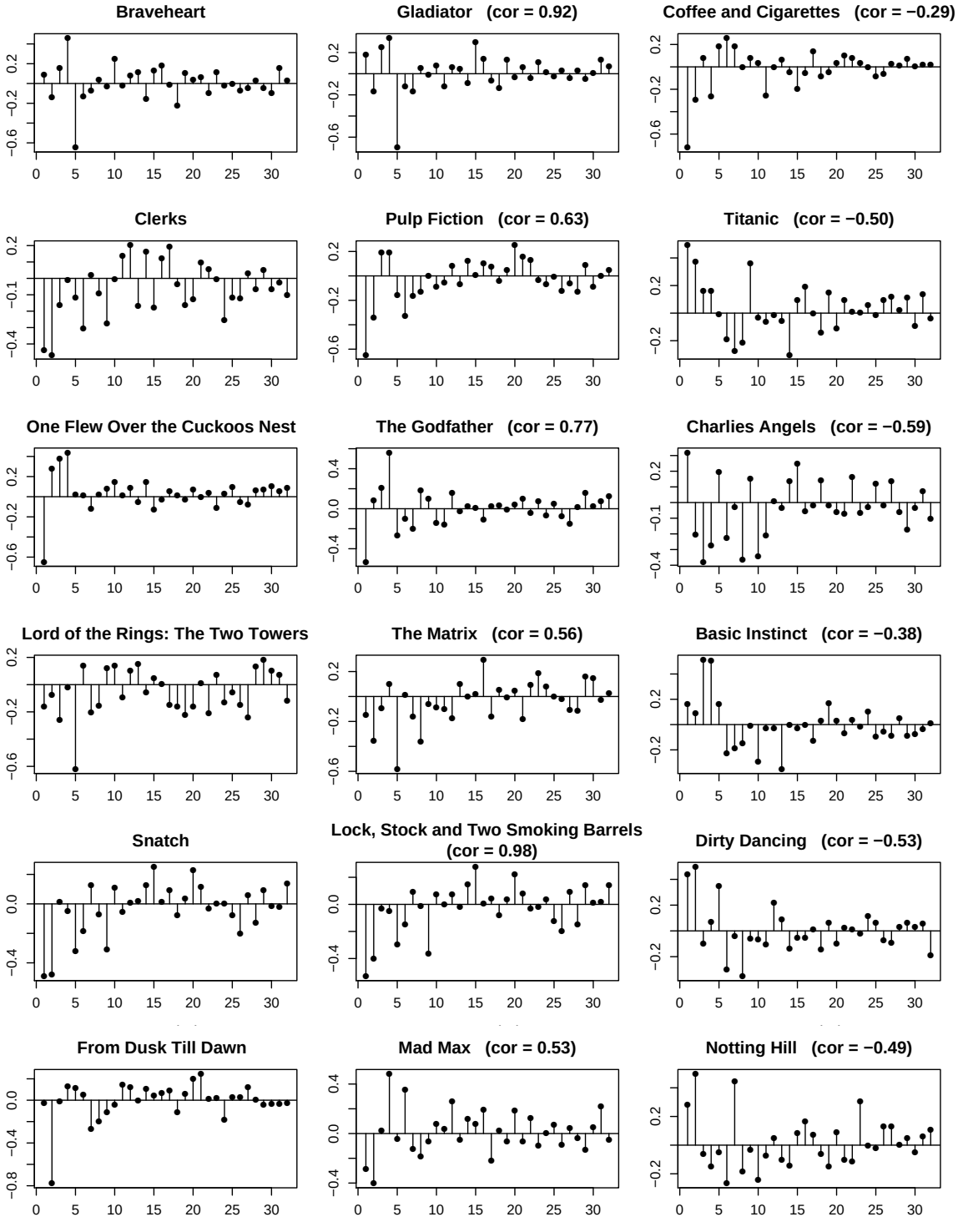

Figure 31: Example movies with their learned feature values, part 2

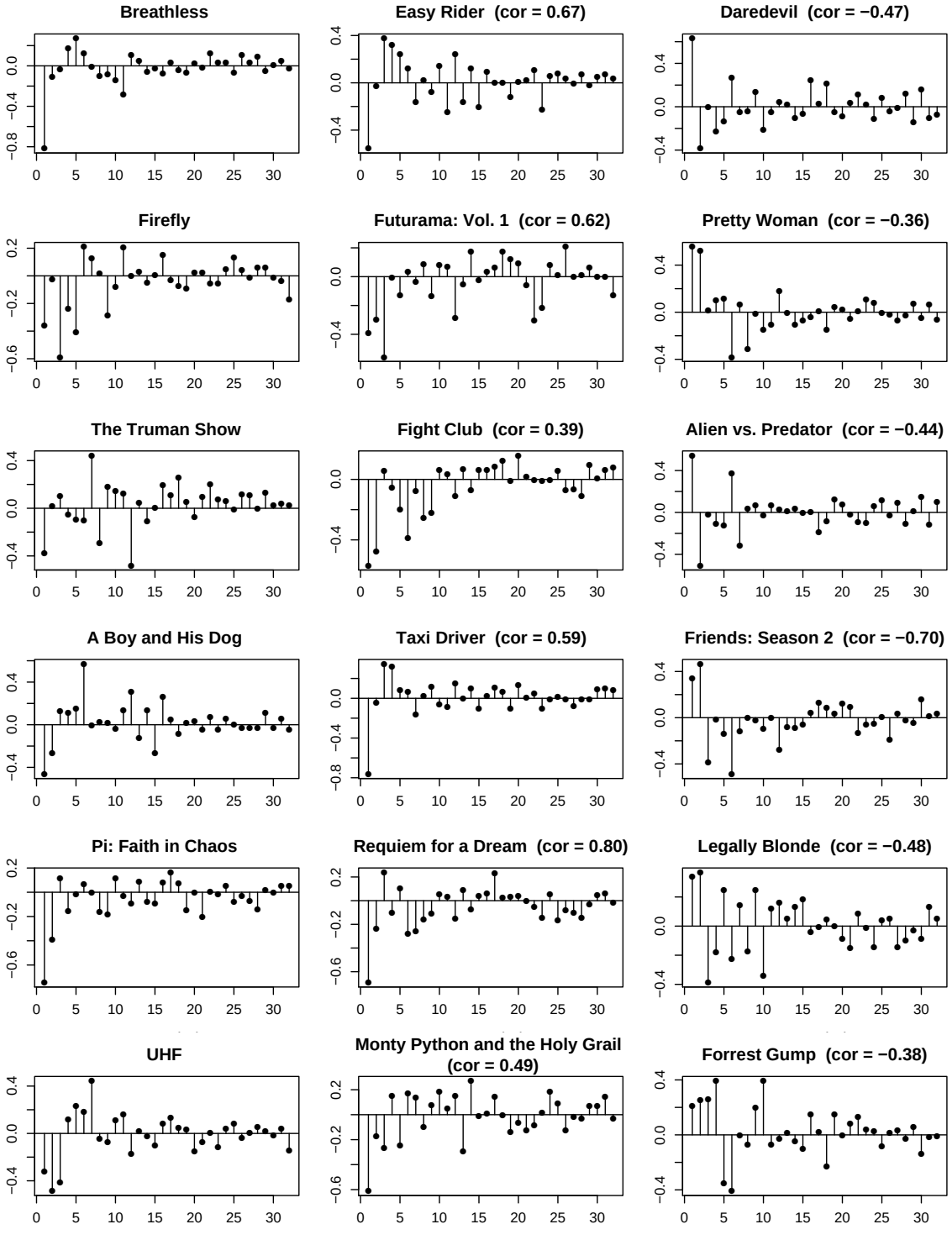

Table 17: Minimal and maximal values of features, among 4000 most popular movies.

Feature 1	Rank	Movie
0.6310	373	Daredevil
0.5550	6	Pretty Woman
0.5450	568	Alien vs. Predator
0.4920	44	Titanic
...	...	...
-0.7210	1731	Coffee and Cigarettes
-0.7450	913	Pi: Faith in Chaos
-0.7630	255	Taxi Driver
-0.8180	2587	Breathless

Feature 4	Rank	Movie
0.5630	130	The Godfather
0.5050	381	Basic Instinct
0.4820	739	Mad Max
0.4590	60	Braveheart
...	...	...
-0.2250	373	Daredevil
-0.2360	1947	Firefly
-0.2640	1731	Coffee and Cigarettes
-0.2750	491	Charlies Angels

Feature 2	Rank	Movie
0.5230	6	Pretty Woman
0.4980	45	Dirty Dancing
0.4970	337	Notting Hill
0.4660	878	Friends: Season 2
...	...	...
-0.4780	333	Snatch
-0.4870	2241	UHF
-0.5080	568	Alien vs. Predator
-0.7780	946	From Dusk Till Dawn

Feature 5	Rank	Movie
0.3480	45	Dirty Dancing
0.2720	2587	Breathless
0.2450	114	Legally Blonde
0.2390	935	Easy Rider
...	...	...
-0.5820	48	The Matrix
-0.6200	30	LotR: The Two Towers
-0.6480	60	Braveheart
-0.6990	32	Gladiator

Feature 3	Rank	Movie
0.5140	381	Basic Instinct
0.3800	227	One Flew Over the (...)
0.3760	935	Easy Rider
0.3470	255	Taxi Driver
...	...	...
-0.3880	878	Friends: Season 2
-0.4140	2241	UHF
-0.5580	1714	Futurama: Vol. 1
-0.5880	1947	Firefly

Feature 6	Rank	Movie
0.5710	3896	A Boy and His Dog
0.3690	568	Alien vs. Predator
0.3570	739	Mad Max
0.2650	373	Daredevil
...	...	...
-0.3810	6	Pretty Woman
-0.3900	85	Fight Club
-0.4090	7	Forrest Gump
-0.4880	878	Friends: Season 2

The next table 18 shows my interpretation of the first six automatically learned features in the regularized SVD, that explain most of the variability of ratings.

Table 18: Interpretation of the first six features.

Feature	Large positive value	Large negative value
1	good defeats evil, idealized world	talking, thinking, emotions, realism
2	love, safety	action, pace, surprise
3	individualism, untrustworthy environment	fairy tale
4	violence, aggression, men's world, patriarchy	independent women, feminism, matriarchy
5	gentle manner, innocence, anxiety	fight, heroism, courage
6	science-fiction, quest, journey	youth, growing up, friendship

We could also try to name the positive and negative part of a feature with single words. My attempt to name the first six features: Idealization vs. Realism, Safety vs. Surprise, Distrust vs. Fairy Tale, Testosterone vs. Feminism, Innocence vs. Heroism, Journey vs. Growing Up.

The above interpretation of features was created based on observation of sorted lists of movie titles. I performed an additional experiment that confirmed to some extent the proposed interpretation – table 19 lists the keywords from IMDb database that are the best predictors of the first six SVD features, according to a chosen criterion. Each feature was predicted by ridge regression with a regularization constant $\lambda = 10$, with predictors

chosen by greedy feature selection among 21 genres and 579 keywords from the IMDb that appear at least 50 times among 2000 movies with the largest support in the Netflix database. The scoring criterion for features in the greedy feature selection was the sum of variance explained and a fraction of the correlation between the IMDb feature and the predicted SVD feature: $\sum_i(\hat{y}_i - \overline{y})^2 / \sum_i(y_i - \overline{y})^2 + 0.05\ |Cor\ (X_k, Y)|$. Each of the six features from the regularized SVD was predicted by 30 IMDb features (genres and keywords) chosen by the greedy feature selection.

Table 19: First six SVD features labeled by IMDb keywords.

Feat.	Var. Expl.	Positive coefficient	Negative coefficient
F1	46%	Action, Family, Thriller, Romance, helicopter, chick-flick, box-office-flop, beautiful-woman	independent-film, cult-favorite, Drama, black-comedy, satire, singing, famous-score, surrealism, Documentary, extramarital-affair, cigarette-smoking, nudity, writer, critically-acclaimed, cafe, adultery, voice-over-narration, drug-use, 1950s, sex, hotel, female-nudity
F2	53%	Drama, Romance, blockbuster, friendship, female-protagonist, Family, male-female-relationship, based-on-novel, 1930s, musician, family-relationships	cult-favorite, Thriller, Action, blood, crude-humor, Horror, Comedy, shot-to-death, topless-female-nudity, sequel, shot-in-the-chest, Sci-Fi, violence, falling-from-height, blood-spatter, gore, masturbation, person-on-fire, punched-in-the-face
F3	53%	Drama, Thriller, murder, sex, neo-noir, adultery, based-on-novel, death, psycho-thriller, gun, Crime, female-nudity, box-office-flop, husband-wife-relationship, beating, shot-in-the-head	Comedy, Family, cult-favorite, Adventure, Animation, martial-arts, Fantasy, Musical, spoof, sequel, lifting-someone-into-the-air, magic, anti-hero, monkey
F4	38%	blockbuster, cult-favorite, famous-score, shootout, famous-line, Sport, revenge, Western, first-of-series, automobile, Crime, racial-slur	flashback, mother-daughter-relationship, Romance, cell-phone, box-office-flop, dancing, Fantasy, tears, female-protagonist, writer, Comedy, love, father-daughter-relationship, drink, homosexual, drinking, painting, graveyard
F5	36%	Comedy, female-protagonist, Horror, teen-angst, Music, Family	Action, no-opening-credits, honor, blockbuster, shot-in-the-forehead, slow-motion, warrior, History, epic, surprise-ending, battle, Adventure, sword-fight, explosion, redemption, cia, monkey, shot-in-the-chest, space-craft, world-war-two, hero, sword, shot-to-death, bravery
F6	38%	Action, Sci-Fi, Adventure, horse, box-office-flop, Thriller, Fantasy, world-war-two, spacecraft, pursuit, alien, murder, Family, future, outer-space	Comedy, boyfriend-girlfriend-relationship, coming-of-age, crude-humor, marijuana, gay-slur, sex, high-school, cocaine, roommate, teen-movie, friendship, obscene-finger-gesture, racial-slur, teen-angst

As we see in table 19, a relatively large amount of variance was explained by binary IMDb tags, although perhaps more tags are needed to describe features 4, 5, 6. The listed values of variance explained were estimated on the training data, and are heightened due to overfitting.

When "folksonomy" tagging is used (tags edited by community, as in IMDb), the quantity and quality of tags increase with movie popularity. For the 2000 most popular movies from the Netflix dataset, as we see from the amount of variance explained, the tags predict the SVD features well. Tags can be used, for example, to produce good priors for item features in SVD-type algorithms, which can be useful in cold start situations for items, often encountered in recommender systems (see the use of IMDb tags to predict features of movies without any ratings in the Netflix database – described in section 4.7). But for movies with more ratings, such as the movies in the Netflix Prize dataset, experiments gave counter-intuitive results, that augmenting the movies with metadata has no effect on accuracy – a small number of ratings is more informative than any amount of metadata [Pil09b, Lee08].

One can try to interpret the automatically learned features from the viewpoints of psychology, anthropology, sociology, culture, ethics. Each movie can tell something about groups of people who rate it: can imply capability of feeling emotions, can tell about personal value systems, moral codes, worldviews, ideological beliefs. The most significant features mark out traits of many users, indicated by many movies. We can surmise that the features depend on gender (features 2+ and 4- likely indicate female, and 4+ male), young age (6+), character traits like sensitivity, neuroticism (feature 1-), or life attitudes like conformism, acceptance, orderliness (feature 1+). We can trace the meaning of features to capability of feeling emotions, like fear or anxiety (feature 5+), or hormone levels heavily influencing emotions, such as high testosterone level (features 4+, 5-) or dopamine (2-, 6+). We could also search for interpretations of the automatically learned information by looking at directions other than the unit vectors in the space of the most meaningful SVD features.

The above interpretations of features, although supported by the observed data, are of course only guesswork, and would require confirmation by conducting appropriate surveys by a trained psychologist.

We can spot in the above examination some connection to movie story types. In the Hollywood Stories dataset, released in 2011, in addition to the standard genres, movies were annotated with the following 22 story types: comedy, love, monster force, quest, rivalry, discovery, pursuit, revenge, transformation, maturation, rescue, escape, the riddle, journey and return, underdog, sacrifice, temptation, fish out of water, metamorphosis, tragedy, wretched excess, rags to riches.

Prediction in the opposite direction is possible – to predict the IMDb tags using the SVD features as predictors, for example, using the logistic regression. Such a prediction can be useful for identifying missing tags, recommending tags in a folksonomy tagging process, or discovering rotations of the feature space – new features most understandable for users. Table 20 shows 50 tags most accurately predicted by the 32 SVD features. Accuracy was measured by the deviance ratio in the logistic regression. The tags tested were a subset of all IMDb tags with support at least 50 among the 2000 most popular movies in the Netflix database.

Table 20: 50 tags best predicted by the SVD features.

chick-flick, crude-humor, spacecraft, teen-movie, future, spoof, warrior, neo-noir, epic, sequel, monster, magic, outer-space, blockbuster, famous-score, psycho-thriller, secret-agent, gothic, gore, based-on-tv-series, robot, sword-fight, alien, surrealism, teen-angst, supernatural-power, battle, blood-spatter, part-of-trilogy, serial-killer, organized-crime, tough-guy, coming-of-age, spy, sword, female-protagonist, impalement, drug-dealing, cult-favorite, shot-in-the-forehead, famous-line, exploding-body, good-versus-evil, castle, time-travel, no-opening-credits, martial-arts, shot-to-death, honor, monkey

Table 21 shows 50 tags (with support at least 50) worst predicted by the SVD features.

Inaccurate prediction indicates that the meaning of those tags is distant from the meaning of the features important for prediction. Those tags are likely needless for predicting ratings or related purposes, and are candidates for removal from the set of tags.

Table 21: 50 tags worst predicted by the SVD features.

title-spoken-by-character, wheelchair, mistaken-identity, dog, bridge, church, funeral, cat, coffee, bus, beach, death-of-friend, drunkenness, reporter, brother-brother-relationship, hospital, airplane, running, father-son-relationship, mirror, one-word-title, california, arrest, friend, christmas, library, interview, jail, punctuation-in-title, birthday-party, kitchen, father-daughter-relationship, fight, train, diner, rivalry, bar, nurse, prologue, bicycle, book, boy, truck, secret, remake, snow, bird, basketball, police-car, manhattan-new-york-city

Now we will look closer at what the SVD features tell us about the standard move genre classification. Going back to explaining SVD features by metadata, table 22 is similar to the previous table 19, but here predictors were only genres. The 6 predictors for each SVD feature were chosen from the 21 standard genres (genre classification by IMDb), as earlier, using greedy feature selection in ridge regression.

Table 22: First six SVD features labeled by standard genres.

Feature	Variance Explained	positive coefficient	negative coefficient
F1	20%	Action, Family, Romance, Sport, Thriller	Drama
F2	33%	Drama, Family, Romance	Action, Comedy, Horror
F3	43%	Drama, Thriller	Action, Animation, Comedy, Family
F4	9%	Action, Crime, Sport, War, Western	Romance
F5	20%	Comedy, Family, Horror	Action, Adventure, History
F6	25%	Action, Family, Sci-Fi, War, Western	Comedy

Some of the six most meaningful SVD features are less well explained than the others by the standard genre labels. This suggests that the standard set of genres needs to be augmented by creating new meaningful genres (bearing in mind, that binary 0-1 genres have limited power of expression, comparing to the SVD features, which can take positive or negative values on a continuous scale). For example, only 9% of variance of the fourth feature was explained by existing genres – the fourth feature was interpreted earlier (table 18) as "violence, aggression, men's world, patriarchy vs. independent women, feminism, matriarchy", or shorter, as "Testosterone vs. Feminism". Based on these guessed meanings, one or two new binary genres could be created to explain better the fourth feature. Features 1, 5, 6 may also suggest new genres. Features 3 and 2 appear to be sufficiently well described by the standard genres. Other linear combinations (other than the unit vectors) of the top SVD features can also be examined and interpreted, to discover important characteristics of movies, unnamed by the standard taxonomies.

One might wonder whether some of the standard genres are unnecessary and can be removed. To find it out, a reverse procedure to the previous was carried out – predicting the binary genres with the logistic regression, using all 32 SVD features as predictors (similar prediction of genres was carried out in [Sel11] on IMDb matched with the MovieLens dataset). The same method was used as in the previous experiment (tables 20, 21), where SVD features were used to predict IMDb keywords. If we assume that all important information about movies is included in the SVD features, the relevant genres should

be predicted well by the SVD features. The genres least accurately predicted by logistic regression, as measured by the ratio of deviance vs. null deviance, were: Adventure, Biography, Crime, Drama, Music, Mystery. Some of these genres are less well predicted, because they are very common, and thus have more "fuzzy" meaning (for example, Drama appears in 1012 movies among the 2000 most popular used in this study). On the basis of the results, in my judgement the candidates to remove are three of the less common genres: Biography, Music (leaving Musical), and Mystery.

To complement the examination of the standard genres, table 23 lays out the average values of first six SVD features for movies (among the 2000 most popular) having the given genre tag. Because the columns 4 and 6 contain only a few, and small negative values, it suggests that the standard genre labelling cannot express the SVD genres 4- (the presumed meaning is "independent women, feminism"), and 6- ("youth, growing up, friendship")

Table 23: Average values of features for standard genres.

Genre (IMDb)	Count	F1	F2	F3	F4	F5	F6
Action	459	0.28	-0.21	-0.03	0.13	-0.07	0.17
Adventure	365	0.16	-0.06	-0.14	0.11	-0.04	0.17
Animation	96	-0.00	0.01	-0.31	0.03	0.01	0.05
Biography	94	-0.25	0.26	0.20	0.03	0.07	0.01
Comedy	826	0.07	-0.03	-0.12	0.00	0.12	-0.05
Crime	407	0.02	-0.12	0.12	0.11	0.02	0.02
Documentary	36	-0.43	0.04	0.00	-0.08	0.02	-0.04
Drama	1012	-0.07	0.12	0.13	0.03	0.05	0.01
Family	206	0.25	0.17	-0.26	0.05	0.12	0.11
Fantasy	233	0.11	-0.08	-0.16	-0.01	0.04	0.12
History	71	-0.10	0.17	0.13	0.12	-0.15	0.15
Horror	112	0.10	-0.38	0.02	0.06	0.19	0.09
Music	128	-0.05	0.17	-0.14	0.01	0.17	0.01
Musical	53	-0.02	0.31	-0.26	0.08	0.16	0.07
Mystery	244	-0.01	-0.07	0.13	0.02	0.01	0.10
Romance	541	0.10	0.19	-0.01	-0.05	0.10	-0.04
Sci-Fi	201	0.11	-0.23	-0.08	0.06	-0.02	0.24
Sport	86	0.17	0.01	0.01	0.19	0.04	-0.08
Thriller	596	0.15	-0.15	0.13	0.09	0.01	0.10
War	75	-0.08	0.10	0.20	0.19	-0.14	0.16
Western	49	0.02	0.01	-0.03	0.26	-0.05	0.19

Table 24 shows the 20 most correlated pairs of genres, and the 20 least correlated pairs. Visible large positive and negative correlations suggest possible inefficiencies and redundancy in the standard set of genres (but not necessarily – correlated directions can have the same expressive power as uncorrelated ones).

One pattern noticed in the feature values is that the average taste changes with popularity. The global average of normalized vectors of item features is (-0.155 -0.032 -0.172 -0.029 0.239 0.212 ...), and the average for the most popular 200 movies is (0.19 0.013 0.057 0.0077 -0.18 -0.11 ...). Various interpretations of the observed pattern are possible: one is that the pattern represents the notion of popular taste. One may wonder here, if the relationship is causal – whether a produced movie has a better chance to become popular when it has features 1+ (idealization), 5- (heroism) and 6- (growing up), and the features to avoid are 1- (realism), 3- (fairy tale), 5+ (innocence) and 6+ (journey). Another interpretation is that the pattern is a result of different inaccuracies in the regularized SVD algorithm appearing for groups of items with largely different amount of ratings, and the automatically learned hidden genres may correct the prediction for those inaccuracies.

Table 24: Largest correlations between average genre vectors.

Genre 1	Genre 2	Cor.
Music	Musical	0.93
History	War	0.92
Biography	Drama	0.86
Crime	Thriller	0.79
Adventure	Fantasy	0.78
Adventure	Sci-Fi	0.76
Action	Thriller	0.76
Action	Adventure	0.75
Action	Sci-Fi	0.75
Fantasy	Sci-Fi	0.73
Mystery	Thriller	0.73
Family	Musical	0.72
Animation	Family	0.72
Family	Fantasy	0.69
Animation	Musical	0.69
Biography	History	0.68
Adventure	Family	0.68
Biography	Documentary	0.65
Animation	Fantasy	0.64
Horror	Thriller	0.62

Genre 1	Genre 2	Cor.
Action	Documentary	-0.64
Action	Biography	-0.63
Comedy	History	-0.58
Comedy	War	-0.58
Music	Thriller	-0.55
Crime	Musical	-0.51
Crime	Music	-0.49
Biography	Sci-Fi	-0.48
Adventure	Documentary	-0.48
Action	Drama	-0.47
Adventure	Biography	-0.47
Biography	Comedy	-0.46
Adventure	Drama	-0.46
Documentary	Thriller	-0.46
Biography	Fantasy	-0.45
Musical	Thriller	-0.45
Drama	Fantasy	-0.45
Crime	Family	-0.43
Comedy	Mystery	-0.41
Drama	Sci-Fi	-0.41

More experiments are needed on varied data to decisively explain this observed pattern.

The above analyses relating SVD features and IMDb genres were heuristic, and should be additionally confirmed by directly modelling the influences of genres and keywords on ratings, by building and training an additional model, instead of utilizing only the SVD results. The notion of "genre" should also be rethought, and more precisely defined. The standard genre theory [Cha97] does not give a precise definition of what is a genre, so I assumed here that genres are named subsets of movies, just like another keywords in the IMDb dataset.

Summarizing, I examined the first six features from regularized SVD, using the Netflix data augmented by IMDb keywords and genres. I listed movies with extreme values of the six features, calculated representations of the features with IMDb keywords and genres, proposed an interpretation of the six features as new genres, and also examined the standard set of genres by IMDb, considering, where it should be augmented by new genres, and which standard genres seem unneeded. The subspace of features learned by an SVD-type algorithm can be spanned by different vectors than the unit vectors – we could choose different rotations of the SVD features that would be more "clean", understandable and intuitive for a human. One might wonder, what causes that humans decide on a taxonomy such as "action", "comedy", "horror" as the dimensions of the space to place movies. It is a question entering the domains of psychology and culture, perhaps related to the concept of meme. Understanding the formation mechanism of such taxonomy could result in dimensionality reduction methods with better explainability, which, as user studies show, is a desirable feature in recommender systems. A similar need to examine, improve or create taxonomies with help of dimensionality reduction methods appears also in other domains, like e.g. for factor analysis methods used in psychometrics.

4.3.5 Improved user preferences

As we mentioned earlier, ratings in the Netflix data are not missing completely at random ($p(Indicators) \neq p(Indicators|Ratings_{observed})$) – users tend to watch and rate movies