BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 07/10 pp72–84
Volume 6, Number 1, March 2023
DOI: 10.26599/BDMA.2022.9020024

# Closed-Form Models of Accuracy Loss due to Subsampling in SVD Collaborative Filtering

Samin Poudel* and Marwan Bikdash

**Abstract:** We postulate and analyze a nonlinear subsampling accuracy loss (SSAL) model based on the root mean square error (RMSE) and two SSAL models based on the mean square error (MSE), suggested by extensive preliminary simulations. The SSAL models predict accuracy loss in terms of subsampling parameters like the fraction of users dropped (FUD) and the fraction of items dropped (FID). We seek to investigate whether the models depend on the characteristics of the dataset in a constant way across datasets when using the SVD collaborative filtering (CF) algorithm. The dataset characteristics considered include various densities of the rating matrix and the numbers of users and items. Extensive simulations and rigorous regression analysis led to empirical symmetrical SSAL models in terms of FID and FUD whose coefficients depend only on the data characteristics. The SSAL models came out to be multi-linear in terms of odds ratios of dropping a user (or an item) vs. not dropping it. Moreover, one MSE deterioration model turned out to be linear in the FID and FUD odds where their interaction term has a zero coefficient. Most importantly, the models are constant in the sense that they are written in closed-form using the considered data characteristics (densities and numbers of users and items). The models are validated through extensive simulations based on 850 synthetically generated primary (pre-subsampling) matrices derived from the 25M MovieLens dataset. Nearly 460 000 subsampled rating matrices were then simulated and subjected to the singular value decomposition (SVD) CF algorithm. Further validation was conducted using the 1M MovieLens and the Yahoo! Music Rating datasets. The models were constant and significant across all 3 datasets.

**Key words:** collaborative filtering; subsampling; accuracy loss models; performance loss; recommendation system; simulation; rating matrix; root mean square error

## 1 Introduction

Collaborative filtering (CF) algorithms are the most widely used and successful recommender algorithms applied to various domains[1,2]. CF methods predict whether a new user might like an item based on the known preferences of a list of users towards a list of items. The core idea behind the CF algorithms is that users who agreed in the past are expected to agree in the future. CF algorithms are most popular because of their ability to perform well with available user ratings only, even in the absence of metadata[3].

The recommendations obtained from CF algorithms are based on available ratings of users toward different items stored in a rating matrix[4]. The rating matrices are usually highly sparse and skewed towards few items and few users, which can adversely affect the performance of CF methods[5,6]. Recent studies have focused on proposing new CF approaches and on improving the performance of existing CF approaches while predicting missing ratings in incomplete high dimensional sparse rating matrices[7–11]. However, there are no studies

• Samin Poudel and Marwan Bikdash are with the Department of Computational Data Science and Engineering, North Carolina A & T State University, Greensboro, NC 27401, USA, E-mail: spoudel@aggies.ncat.edu; bikdash@ncat.edu.
* To whom correspondence should be addressed.
Manuscript received: 2022-02-13; revised: 2022-07-09; accepted: 2022-07-11

proposing constant predictive models that predict the performance of a CF method based on the dataset properties only.

It is intuitive to expect that the performance of a CF algorithm depends on the characteristics of a rating matrix like the number of users, number of items, and the density[12], and that the performance deterioration due to subsampling of items and users depends on the level of subsampling. In Ref. [13], a detailed analysis of the effects of the rating matrix characteristics on the performance of CF algorithms was presented. In Ref. [14], the influence of similarity metrics on the behavior of CF methods was studied, and similarity metrics were shown to depend on the available number of users and items in the rating matrix[15]. The study in Ref. [16] showed that the performance of many CF algorithms is directly related to the density of the rating matrix. The authors in Ref. [17] showed that increasing the density of information through advanced imputation prior to applying singular value decomposition (SVD) CF improves its performance.

The influence of rating data characteristics on the performance of CF methods is mainly based on explanatory analysis[12,18,19] and qualitative models[20]. The performance of a CF algorithm was shown in Ref. [21] to degrade when decreasing the number of ratings, and in Ref. [22] it was shown that the performance improves with larger data sets. In Refs. [13,20,23], the authors subsampled the rating data many times, built linear regression models for the performance of CF methods versus the rating data characteristics and used the models for a qualitative explanatory analysis[24,25]. The analysis was qualitative in the sense that the coefficients of the model were specific for every case (combination of dataset, algorithm, and levels of sampling), and hence no quantitative analysis can be provided for an arbitrary case. Similar conclusions were reached in Refs. [26, 27] which can predict the performance of a CF method after subsampling based on the rating matrix characteristics. One often develops a linear regression model of the accuracy loss for a given dataset and a given CF[20], but these models are assumed to be unrelated to each other and to depend on the dataset characteristics in an unknown way. In other words, there is no established constant predictive model.

Our previous investigation[28] showed that the improvement in the training time due to subsampling for a CF algorithm can be modeled using the same linear regression model for a wide range of datasets, algorithms, and levels of subsampling. In fact, the model is exceedingly simple and general and expresses the improvement as

$$T^S/T^P = 1 - \mu - \nu + \mu\nu \tag{1}$$

where $T^P$ is the time to train a CF method using primary rating matrix $P$, $T^S$ is the time to train a CF method using a subsampled rating matrix $S$ from $P$, $\mu$ is the fraction of users dropped (FUD) from $P$, and $\nu$ is the fraction of items dropped (FID) from $P$ to get $S$, defined as

$$\mu = \frac{\text{number of users in } P - \text{number of users in } S}{\text{number of users in } P} \tag{2}$$

$$\nu = \frac{\text{number of items in } P - \text{number of items in } S}{\text{number of items in } P} \tag{3}$$

For example, assume that the number of users in $P$ is 10 000 and the desired number of users in $S$ is 5000. Then the FUD can be computed using Eq. (2) as $\mu = (10\,000 - 5000)/10\,000 = 0.5$. Similarly, if the desired number of items in $S$ is 5000, then the FID can be computed using Eq. (3) as $\nu = (10\,000 - 5000)/10\,000 = 0.5$.

Since all the coefficients in Eq. (1) can be approximated significantly by $\pm 1$, the model can be considered general, constant, and closed-form, thus providing a quantitative analysis of the effect of subsampling on the training time. For example, if half of the items are dropped and half of the users are dropped, then according to Eq. (1), $T^S/T^P = 1/4$, implying that the training time has been reduced 4 times. If 90 percent of the items and users are dropped, then $T^S/T^P = 0.01 = 1/100$, and the training time has been reduced 100 times.

The attempt in Ref. [28] to find an equally simple and general model for the deterioration in the root mean square error (RMSE) due to subsampling was not successful. The investigation revealed however that the deterioration in RMSE due to subsampling is modeled better using the FUD odds ratio $\mu/(1-\mu)$ and the FID odds ratio $\nu/(1-\nu)$ and that the relationship appears to be multi-linear in the odds. The coefficients however presented no simple patterns in terms of the combination of dataset and algorithm.

In this work, we seek to establish a general constant closed-form model for the deterioration in accuracy due to subsampling that is similar to Eq. (1), in the sense that the models are constant across the datasets of different

characteristics and its parameters are determined in terms of the dataset characteristics. In addition to the understanding and intuition that these constant models provide, they also suggest that a theory of subsampling effects is also feasible. Our approach is to construct a constant subsampling accuracy loss (SSAL) model by synthesizing 850 primary rating matrices from the 25M MovieLens dataset[29,30], to subsample each of these primary matrices many times, to recognize the patterns in the subsample CF accuracy loss, and then to abstract a general constant model whose coefficients have simple closed-form expressions in terms of the dataset characteristics. Finally, we will show that the general constant model explains the majority of the variation in the performance across datasets. We considered 3 expressions of accuracy loss, one in terms of the RMSE, on which most of the exploration was performed, and 2 expressions based on the mean square error (MSE). One of the MSE-based expression of accuracy loss was found to be superior.

The primary contributions of this study are listed below.

- This work pioneers the idea of developing constant models of the accuracy of CF predictive algorithms that depend only on the simple properties of the dataset, namely, the dimensions and the densities of the dataset.
- In particular, we developed and validated quantitative SSAL models for SVD CF method that are constant across the datasets. For instance, if 80 percent of the users and items are dropped from a dataset having an average of 20 ratings per user and 20 ratings per item, our best model expects that $\mathrm{MSE}^S/\mathrm{MSE}^P = 7.7$ for any reasonable dataset. Alternatively, we say that the MSE of the SVD CF is increased 7.7 times (see Section **??** for detailed illustration).
- This work and the analysis in Ref. [28] together suggest that constant models can also be achieved for other CF methods as well. The feasibility of a theory of subsampling effects on the performance of CF methods is supported by this work.

## 2 Methodology

### 2.1 Notation and measures of accuracy loss

Let $P$ denote the primary rating matrix (PRM) and $S$ denote the subsampled rating matrix (SRM) from $P$. Let $m$ represent the number of rows and hence the number of users, and let $n$ represent the number of columns, and hence the number of items. Let $\mu$ represent the FUD during subsampling, and $\nu$ represent the FID during subsampling of the PRM.

If one interprets $\mu$ as the probability of dropping a user, then one can define the odds ratio of dropping a user (ORDU) as

$$O_\mu = \frac{\mu}{1-\mu} \tag{4}$$

Similarly, if one interprets $\nu$ as the probability of dropping an item, then one can define the odds ratio of dropping an item (ORDI) as

$$O_\nu = \frac{\nu}{1-\nu} \tag{5}$$

We note that our previous investigation has suggested that the odds are more expressive of the RMSE deterioration due to subsampling[28].

Let $\delta$ denote the density of the rating matrix, which is the number of nonzero elements over the total number of elements $mn$. If $N_r$ denotes the total number of ratings or nonzero elements in a rating matrix. Then

$$\delta = \frac{N_r}{m \times n} \tag{6}$$

and hence $\delta$ can be thought of as a surface density.

The square root of $\delta$ or $\delta^{1/2}$ therefore represents a matrix-level linear density. One can define alternative "linear" densities as follows:

$$\delta_U = \frac{N_r}{m} = n\delta \quad \text{and} \quad \delta_I = \frac{N_r}{n} = m\delta \tag{7}$$

The ratio of linear densities is equal to the ratio of number of users to items or vice versa

$$\frac{\delta_I}{\delta_U} = \frac{m}{n} \tag{8}$$

Here, we interpret $\delta_U$ as the average number of ratings per user and $\delta_I$ as the average number of ratings per item.

A CF application process consists of dividing a rating matrix $R$ into training and test data, learning a model from the training data and testing the performance of the learned model using the test data. In this study, we have used the SVD CF[31] algorithm because of its predictive ability even with datasets having low density and disproportionate numbers of users and items. The SVD CF algorithm was implemented using the SURPRISE python library[32].

The performance of CF methods can be evaluated using a variety of predictive accuracy metrics, classification accuracy metrics, ranking accuracy metrics, and others[33–35]. Various recommender systems perform differently according to the evaluation metrics[36]. The most popular and commonly used metrics for predictive accuracy are the RMSE[16,37] and the mean absolute error (MAE)[34,38]. In this work, we will use the RMSE, or its square, the MSE if advantageous.

The MSE is defined as the average squared error between the observed and predicted values of a variable[39]. Let $r$ be an observed rating in a rating matrix $R$ and $\hat{r}$ be the predicted rating using a CF model, then MSE of the CF model based on $|K|$ number of predictions can be computed as

$$\text{MSE} = \frac{1}{|K|}\sum_{k\in K}(r_k - \hat{r}_k)^2 \tag{9}$$

and the RMSE is its square root.

Let $\text{RMSE}^P$ denote the RMSE achieved by applying a CF technique to a PRM $P$, and $\text{RMSE}^S$ denote the same for a subsampled rating matrix $S$ from $P$. Since the subsampling is generally known to increase the RMSE error, we can write

$$\rho = \frac{\text{RMSE}^S}{\text{RMSE}^P} = 1 + \text{Accuracy Loss} \geqslant 1 \tag{10}$$

Our preliminary investigation[28] has suggested an SSAL model based on RMSE given by

$$\rho = 1 + \eta_0\left(\frac{O_\mu}{1-\nu}\right) + \eta_1\left(\frac{O_\nu}{1-\mu}\right) - \eta_2 O_\mu O_\nu \tag{11}$$

where the $\eta$'s are regression coefficients that are determined for each primary rating matrix by constructing many (say 550) subsampled rating matrices and computing the RMSE for each subsampled rating matrix.

The model in Eq. (11) is obviously dependent on the primary matrix $P$, and $P$ itself can be characterized by its density $\delta(P)$ and its dimensions ($m(P)$ and $n(P)$, the numbers of rows and columns), and other primary matrix characteristics (PMC). The objective of this paper is to study the dependence of these models (and their coefficients) on the PMCs. If the dependencies are simple, then a general constant closed-form SSAL would be derived.

### 2.2 Overview of the methodology

Here, we start with the 25M MovieLens dataset[29,30] ($\text{M} = 10^6$, reflecting the approximate number of ratings in each dataset) and synthesize about 850 distinct Primary Rating Matrices from it using judgmental sampling. Care is taken to generate primary matrices with a wide and representative range of $m$, $n$, and $\delta$, and perhaps other PMCs. For each PRM, nearly 550 subsampled rating matrices are synthesized using density-constrained subsampling algorithm proposed in Ref. [28]. Each SRM $S$ has a unique combination of fractions of users and items dropped, $\mu$ and $\nu$. For each SRM, the accuracy measure of the SVD CF model is computed. For each PRM, linear regression is used to compute the coefficients in the SSAL(RMSE) model in Eq. (11).

Subsequently, we perform various visualization tests in an attempt to ascertain whether the coefficients have a simple dependency on the PMCs. This includes, for instance, plotting a coefficient vs. a PMC or vs. another coefficient. Then, one SSAL(RMSE) model that predicts the accuracy loss for all considered primary matrices was derived following the 3-step procedure explained in Section 4.4. Similarly, we modeled SSAL based on MSE and this lead to a more elegant SSAL model in Eq. (25).

## 3 Generating the Primary Matrices

We will use the 1M MovieLens (1M-ML) dataset[30,40], the 25M MovieLens (25M-ML) dataset[29,30] and, the Yahoo! Music (YM) dataset[41] during this study. The steps involved in extracting a PRM $P$ having $m$ users and $n$ items and a density $\delta$ are as follows.

(1) Update the original rating dataset by dropping rows or columns or both using judgmental sampling until it has the density $\delta$. One can drop highly dense rows/columns if $\delta$ is less than the density of original dataset. Similarly, less dense rows/columns can be dropped if $\delta$ is higher than the density of original dataset from which PRM is being extracted. Care must be taken that the number of users does not become less than $m$ and the number of items does not become less than $n$ in the updated original dataset, as sampling with replacement has not been considered while extracting a PRM.

(2) Find the fraction of users that has to be dropped from the updated original dataset in Step 1 so that it can be updated to a rating matrix of $m$ users. Also, compute the fraction of items to be dropped from the updated original data set that leads to a rating matrix with $n$ users.

(3) Apply the density-based random stratified subsampling using clustering (DRSC) proposed in Ref. [28] to the updated original dataset in Step 1 using the values of FID and FUD from Step 2. The resulting rating matrix is the PRM $P$ with the desired number of users $m$, number of items $n$, and density $\delta$.

The primary rating matrices extracted from 1M MovieLens dataset have rating data in discrete numerical rating scale of 1 to 5. Primary rating matrices extracted from 25M MovieLens dataset[29] have rating data in discrete numerical rating scale of 0.5 to 5 in steps of 0.5. Primary rating matrices extracted from Yahoo! music dataset[41] have rating data in discrete numerical rating scale of 1 to 100.

Primary rating matrices considered to postulate the

preliminary SSAL models are extracted from the 25M MovieLens dataset[29,30]. A representative number of PRMs is shown in Table 1. For every size of synthesized PRM, we have varied density as mentioned in Table 1. Around 850 primary matrices differing in at least one of the dataset characteristics were synthesized. From each PRM, around 550 density constrained subsampled rating matrices were obtained. Hence, our analysis of results involves application of SVD CF to nearly 460 000 subsampled rating matrices.

## 4 Analysis of the Coefficients of Subsampling Accuracy Loss Model Based on RMSE

### 4.1 Variation of coefficients of SSAL model with primary rating matrix characteristics

For each primary matrix $P$ with parameters $m(P)$, $n(P)$, and $\delta(P)$, we have a regression model of the SSAL(RMSE) with parameters $\eta_0(P)$, $\eta_1(P)$, and $\eta_2(P)$. In the following, we will drop the explicit dependence on $P$ and discuss the observed relationships between the above 6 quantities that are function of $P$.

In the first (discovery) phase, we will try to discover the likely relationships, by plotting the coefficients for various special cases. Figure 1 suggests that for given $m = 4000$ and $n = 5000$, the coefficients $\eta_i$ depend weakly on the density of the primary matrix. Moreover, $\eta_2$ appears to be the sum of $\eta_0$ and $\eta_1$.

Qualitatively similar conclusions were reached by considering other combinations of $m$ and $n$. A case which showed a slightly wilder variation is more unbalanced with ($m = 9000$, $n = 3000$), and it is shown in Fig. 2. But even so, the above relationships seem to hold.

Based on the Figs. 1 and 2, we can state the following conclusions.

(1) A change in the density of PRM does not affect the values of $\eta's$ significantly at a constant size of the users and items of PRM.

**Table 1 Details of synthesized primary rating matrices from 25M MovieLens dataset.**

| Number of users ($m$) in $P$ | Number of items ($n$) in $P$ | Density of $P$ | Number of primary matrices |
|---|---|---|---|
| 7000 | 5000 | 0.06, 0.08, ..., 0.18 | 7 |
| 8000 | 4200, 4500, ..., 12 000 | 0.06, 0.1 | 52 |
| 6000 | 3000, 3300, ..., 9000 | 0.06, 0.08, 0.09, 0.11, 0.12, 0.14 | 120 |
| 4200, 4500, ..., 12 000 | 8000 | 0.06 | 26 |
| 3000, 3300, ..., 9000 | 6000 | 0.06, 0.08, 0.09, 0.11, 0.12, 0.14 | 120 |
| 3000, 4000, 5000 | 3000, 4000, 5000 | 0.05, 0.06, ..., 0.2 | 180 |
| 5000 | 3000, 3500, ..., 10 000 | 0.09, 0.1 | 28 |
| 5000 | 6000, 7000, ..., 10 000 | 0.05, 0.06, ..., 0.2 | 100 |
| 3000, 3500, ..., 10 000 | 5000 | 0.09, 0.1 | 28 |
| 6000, 7000, ..., 10 000 | 5000 | 0.05, 0.06, ..., 0.2 | 100 |
| 8000 | 12 000 | 0.05, 0.06, ..., 0.2 | 20 |
| 9000, 12 000 | 3000 | 0.05, 0.06, ..., 0.2 | 40 |
| 10 000 | 10 000 | 0.05, 0.06, ..., 0.2 | 20 |

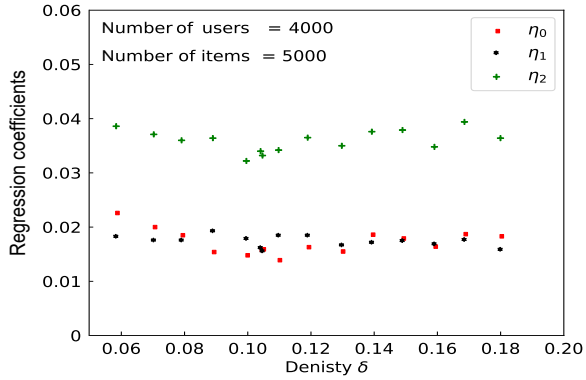


**Fig. 1 Analysis of coefficients of SSAL model versus density at constant number of users and items for $P$ (4000, 5000).**

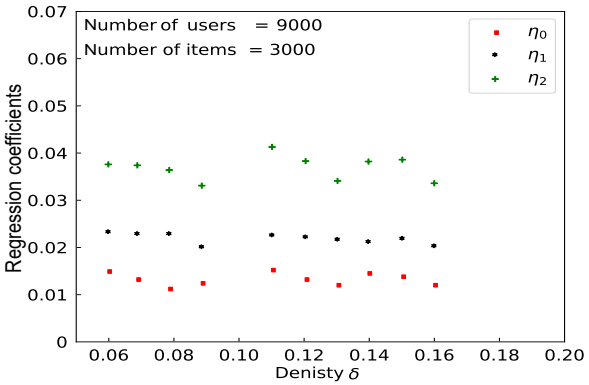


**Fig. 2 Analysis of coefficients of SSAL model versus density at constant number of users and items for $P$ (9000, 3000).**

(2) The hypothesis $\eta_2 \approx \eta_0 + \eta_1$ relating the coefficients of SSAL model is reasonable and should be tested.

Figure 3 depicts the changes in $\eta_i$ with the number of users $m$ in $P$ at constant density and number of items $n$ for $n = 6000$ and $\delta = 0.06$. Other combinations (not shown) exhibited a qualitatively similar behavior which can be summarized as follows.

(1) $\eta_0$ decreases linearly with $m$, the number of users in the PRM at constant density and number of items.

(2) $\eta_1$ increases linearly with $m$ at constant density and number of items.

(3) $\eta_2$ is nearly constant when changing the number of users of PRM at the constant number of items and density.

(4) The hypothesis $\eta_2 \approx \eta_0 + \eta_1$ based on Figs. 1 and 2 and proposed above, seems to be supported by Fig. 3 as well.

Similarly, Fig. 4 shows the variation in $\eta_i$ with the number of items in the PRM at the constant number of users ($m = 6000$) and constant density $\delta = 0.14$. Other combinations were also tested and they showed a qualitatively similar behavior.

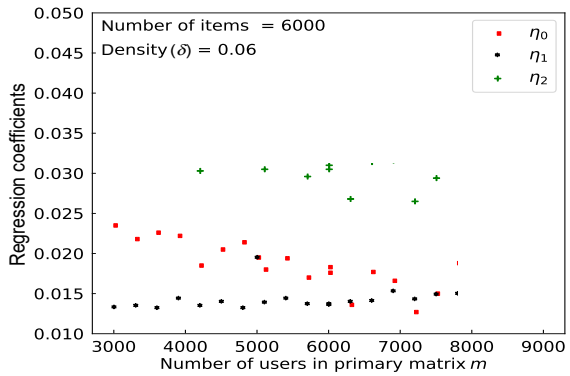


**Fig. 3 Analysis of coefficients of SSAL model versus number of users at constant density of 0.06 and 6000 items.**

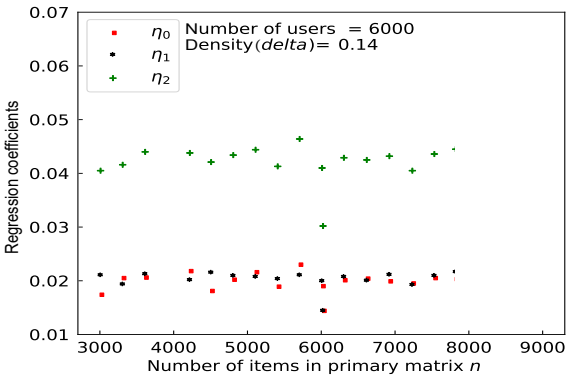


**Fig. 4 Analysis of coefficients of SSAL model versus number of items at constant density of 0.14 and 6000 users.**

From Fig. 4 the following conclusions are suggested.

(1) $\eta_0 \approx \eta_1$ regardless of the number of items in the PRM at a constant number of users and a constant density.

(2) $\eta_0 = \theta_1$ for $m < \kappa n$ and $\eta_0 = \theta_2$ for $m \geqslant \kappa n$ at the constant number of users and constant density, where $\theta_1$ and $\theta_2$ are two values of $\eta_0$ that must be computed from the data. From our analysis, $\kappa \approx 1.35$. In other words, if the number of users is less than 1.35, the number of items one model is needed, and if it is larger, a slightly different model is needed. In the following development, we ignore this fact, which we consider a secondary effect, and seek a model that covers all cases in Table 1.

(3) $\eta_2$ is nearly constant with when changing the number of items of PRM at the constant size of the users and the constant density.

(4) The hypothesis $\eta_2 \approx \eta_0 + \eta_1$ supported by Figs. 1–3 is seem to be true in Fig. 4 as well.

Based on the plots of $\eta$'s versus the different PRM characteristics, we are able to propose the hypothesis

$$\eta_2 \approx \eta_0 + \eta_1 \tag{12}$$

The plots seem to suggest that $\eta_1$ is proportional to $\eta_0$. We assume that the proportionality constant

$$\alpha \triangleq \eta_1/\eta_0 \tag{13}$$

depends on the PRM characteristics. If this is true, then using Formula (12) in Eq. (11) simplifies the RMSE SSAL model to

$$\rho - 1 = \eta_0 O_\mu + \eta_1 O_\nu \tag{14}$$

The notation $\eta_1 = \alpha\eta_0$ will lead to

$$\rho - 1 = \eta_0 \left(O_\mu + \alpha O_\nu\right) \tag{15}$$

The hypothesis proposed in Formula (12) is tested in Section 4.2. The relationship of $\alpha = \eta_1/\eta_0$ with the number of the users and items of PRM is studied in the Section 4.3.

### 4.2 Testing the hypothesis on the relationship among coefficients of SSAL model

We plotted $\eta_0 + \eta_1$ versus $\eta_2$ to test the hypothesis in Formula (12), and the result is shown in Fig. 5. A linear regression analysis is then performed for the foumula.

$$\eta_0 + \eta_1 \approx \text{slope } \eta_2 + \text{intercept} \tag{16}$$

The least-square linear regression yields a slope = 0.96 and the intercept = 0.0018, which supports the proposed hypothesis of $\eta_0 + \eta_1 = \eta_2$. The $p$-value is 0.00 for the linear regression analysis. The mean absolute error (MAE) during the linear regression analysis is 0.0005. The relative error is therefore about $0.0005/0.05 = 1\%$ or about $0.0005/0.02 = 2.5\%$ in the worst case.

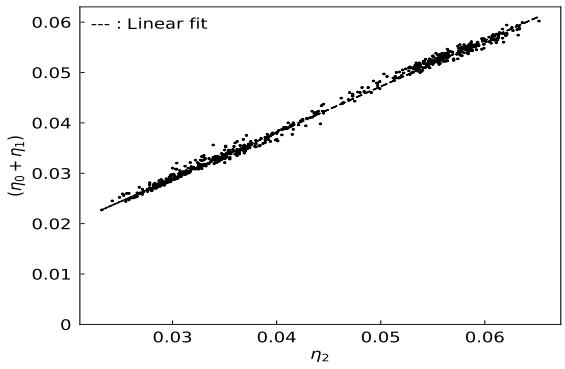


**Fig. 5 Analysis of coefficients of SSAL model.**

### 4.3 Analysis of the ratio of coefficients of SSAL model with PMCs

Figure 6 shows the variation of $\alpha = \eta_1/\eta_0$ with the number of users at constant number of items $n = 6000$. The relationship appears to be linear. Other combinations showed qualitatively the same patterns.

Figure 6 shows that $\alpha$ increases linearly with the number of users of PRM at the constant number of items. Also, Fig. 6 shows that $\alpha$ seems to be higher for higher value of density for a constant size of the users and items. Figure 7 shows the dependence of $\alpha = \eta_1/\eta_0$ on the the number of items of PRM for 6000 users. Other

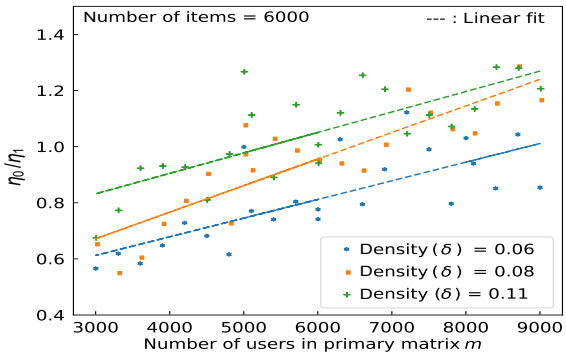


**Fig. 6 Analysis of $\alpha = \eta_1/\eta_0$ versus number of users at constant 6000 items.**

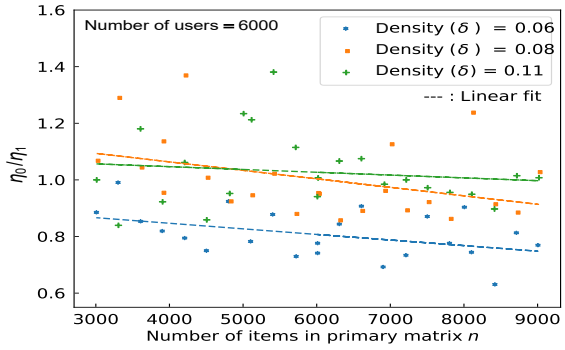


**Fig. 7 Analysis of $\alpha = \eta_1/\eta_0$ versus number of items at constant 6000 users.**

simulations (not included in this paper) show the same qualitative behavior.

Figure 7 depicts that $\alpha$ decreases linearly when increasing the number of items of PRM at a constant number of the users. Also, Fig. 7 shows that $\alpha$ seems to be higher for higher value of density for a particular number of users and items. Hence, the assumption that $\alpha$ can be represented in terms of PRM characteristics is validated.

In conclusion, $\alpha$ increases with the number of users (at constant $n$ and $\delta$), decreases with number of items (at constant $m$ and $\delta$), and increases with density (at constant $m$ and $n$). Ultimately this leads to the form in Eq. (21).

### 4.4 Coefficients in terms of linear densities

Model in Eq. (14) is linear in $O_\mu$ and $O_\nu$. Now we will check whether the addition of a cross term $O_\mu O_\nu$ gives a better fit. Therefore, we postulate a model M1 to estimate the ratio $\rho$ as

$$\rho = \phi_0 + \phi_1 O_\mu + \phi_2 O_\nu + \phi_3 O_\mu O_\nu \qquad (17)$$

The extensive regression analysis for the model in Eq. (17) suggested $\phi_0 \approx 1$ and $\phi_3 \approx 0$. The results are statistically significant because the regression analysis based on the model in Eq. (17) yielded a high $R^2 = 0.95$ and low MAE $= 0.009$ in average for 850 primary rating matrices considered in Table 1.

Next, we experimented with several multiple linear regression parameters[42] of $\phi_1$ and $\phi_2$ in terms of different combination of $m, n, \delta$. During the regression analysis process, we explored that the variation in the coefficients of SSAL models are significantly explained by the linear densities $\delta_U$ and $\delta_I$. Therefore, for the rest of this paper, we propose and recommend the following procedure.

**Step 1.** Expand the two coefficients (here $\phi_1$ and $\phi_2$) in a "parallel combination" of the linear densities, in other words, as a linear combination of the inverses of the linear densities.

$$\phi_1 \approx a_0 + \frac{a_1}{\delta_U} + \frac{a_2}{\delta_I} \quad \text{and} \quad \phi_2 \approx b_0 + \frac{b_1}{\delta_U} + \frac{b_2}{\delta_I} \qquad (18)$$

where $a$ and $b$ are numerical constants. Moreover, $a_0$ and $b_0$ are the intercepts.

**Step 2.** Evaluate the goodness of fit for the proposed regression models for the coefficients of SSAL models.

The regression models based on Eq. (18) are shown in Table 2. The coefficient of determination $R^2$ illustrates that about 80 percent of the variations in the coefficients of SSAL(RMSE) model in Eq. (15) can be explained by

**Table 2 Regression coefficients based on regression analysis of $\phi_1$ and $\phi_2$ versus the linear densities, $\delta_U$ and $\delta_I$ in Eq. (18).**

| Regression of | Coefficient of $(1/\delta_U)$ | Coefficient of $(1/\delta_I)$ | Intercept | $R^2$ |
|---|---|---|---|---|
| $\phi_1$ | $2.00''$ | $4.38'$ | 0.0048 | 0.84 |
| $\phi_2$ | $4.33'$ | $2.01''$ | 0.0061 | 0.78 |

Note: The double primes $''$ in Table 2 represent a standard deviation of $\sigma \leqslant 0.5$ in coefficients ($a$ and $b$) of Formula (18) as obtained from the 50 fits. A single prime $'$ indicates $0.5 < \sigma \leqslant 0.8$. The standard deviations in the Intercept in Table 2 were less than 0.005 for all cases. All coefficients and intercept in Table 2 had very small $p$-value $p \leqslant 0.01$. The $R^2$[43] column is used for evaluation of the regression fit.

the linear densities $\delta_U$ and $\delta_I$.

**Step 3.** Find simple ratios of the coefficients of linear densities.

From the Table 2 and considering the standard deviation in $a$ and $b$ we can postulate that

$$a_1 \approx b_2,\ a_2 = b_1 \ \text{ and } \ \frac{a_2}{a_1} = \frac{b_1}{b_2} \approx 2 \qquad (19)$$

Equation (19) states that $\phi_1 = \phi_2$ when the linear densities are equal.

Hence, using Eqs. (19) and (18) and Table 2, the model M1 in Eq. (17) can be simplified to

$$\rho = 1 + \Phi\left(\frac{\delta_I + 2\delta_U}{\delta_U\delta_I}O_\mu + \frac{2\delta_I + \delta_U}{\delta_U\delta_I}O_\nu\right) = 1 + \Phi\left(\frac{m+2n}{mn\delta}O_\mu + \frac{2m+n}{mn\delta}O_\nu\right) \qquad (20)$$

where $\Phi = a_1 = b_2$. The accuracy loss models in Eqs. (20) are symmetrical because exchanging the values of $(\mu, n)$ with $(\nu, m)$ results in the same SSAL.

The models in Eq. (20) are constant over wide range of PRM data characteristics including different densities, covering the two cases $m > n$ and $m \leqslant n$. For no items dropped, or $O_\nu = 0$, the models suggest that SSAL is more inversely proportional to the average number of ratings per item than average number of ratings per user for a specified $\mu$. Similarly, for no users dropped, or $O_\mu = 0$, the second term in the models suggest that SSAL is more inversely proportional to the average number of ratings per user than the average number of ratings per item for a specified $\nu$. Also, the SSAL is more sensitive to FUD and FID compared to the PRM characteristics.

Based on Table 2, $\Phi$ ranges form 1.5 to 2.5. So, we compared the actual $\phi_1$ and $\phi_2$ for the primary matrices with $\phi_1$ and $\phi_2$ computed from using relation in Eq. (19) for the possible range of $\Phi = a_1 \in [1.5, 2.5]$. Our analysis yielded that $\Phi \approx 2.3$ has the best fits. We will further explore the best value of $\Phi$ for the final model M1 in Eq. (20) for different cases in Section 6.

Also $\alpha = \eta_1/\eta_0$ can be written as

$$\alpha = \frac{\eta_1}{\eta_0} = \frac{\phi_2}{\phi_1} = \frac{2\delta_I + \delta_U}{\delta_I + 2\delta_U} = \frac{2m+n}{m+2n} \qquad (21)$$

## 5 Effect of Subsampling on the MSE

### 5.1 Model with a presumed cross term

If the FID and FUD are very small, perturbation theory suggests that the SSAL model based on the MSE has the same model as the SSAL based on RMSE with an additional factor of 2 multiplying the coefficients $\phi_1$ and $\phi_2$. The question we attempt to answer now is whether an equally simple multi-linear model of the deterioration of the MSE leads to a better approximation over a wider range of FID and FUD; for instance, at extreme sampling levels characterized by very high subsampling rates. Therefore we postulate the model M2 for the ratio $\rho^2 = \dfrac{\text{MSE}^S}{\text{MSE}^P}$ as

$$\rho^2 = \psi_0 + \psi_1 O_\mu + \psi_2 O_\nu + \psi_3 O_\mu O_\nu \qquad (22)$$

Preliminary regression fits suggested that $\psi_0 \approx 1$ and $\psi_3 \approx 0$. The results are statistically significant because the regression analysis based on the model in Eq. (22) yielded high $R^2 = 0.96$ and low MAE $= 0.014$ on average for 850 primary rating matrices considered in Table 1.

The fact that $\psi_3 \approx 0$ and hence the MSE has no cross terms involving $O_\mu O_\nu$ over a wide range of PMC and subsampling levels are very important. It shows that the RMSE model introduces spurious behavior. This is because squaring the RMSE model would lead to a nontrivial cross term, which is contradicted by the evidence.

We apply the 3-step procedure explained in Section 4.4 to the coefficients of the M1 model. In Step 1, we express the coefficients as parallel combination of the linear densities

$$\psi_1 \approx c_0 + \frac{c_1}{\delta_U} + \frac{c_2}{\delta_I},\ \psi_2 \approx d_0 + \frac{d_1}{\delta_U} + \frac{d_2}{\delta_I} \qquad (23)$$

where $c_0$ and $d_0$ are intercepts. Next, Steps 2 and 3 lead to Table 3 which shows that

$$c_1 \approx d_2,\ c_2 = d_1, \ \text{ and } \ \frac{c_2}{c_1} = \frac{d_1}{d_2} \approx 2 \qquad (24)$$

thus leading to the model with the same pattern as before

$$\rho^2 = 1 + \Psi\left(\frac{\delta_I + 2\delta_U}{\delta_U\delta_I}O_\mu + \frac{2\delta_I + \delta_U}{\delta_U\delta_I}O_\nu\right) \qquad (25)$$

where $\Psi = c_1 = d_2$. M2 is a symmetrical model like M1.

The values of $c_1$ , $d_2$ and the relation $c_1 \approx d_2$ in Table 3 suggest that $c_1 = \Psi$ belongs to the range [5.0, 7.0]. We found that $\Psi \approx 5.6$ yields the best fit for models in Eq. (23).

Hence, the expressions for SSAL from MSE and RMSE are qualitatively similar and are related to each other with relation between $\Phi$ and $\Psi$. Meaning,

$$\frac{\rho-1}{\rho^2-1} = \frac{\Phi}{\Psi} \Longrightarrow \text{SSAL(RMSE)} = \tau\text{SSAL(MSE)} \quad (26)$$

### 5.2 Factored MSE model

Finally, we also considered a multi-linear model M3 of the MSE deterioration that has an outer product form.

$$\rho^2 = (1 + \gamma_1 O_\mu)\,(1 + \gamma_2 O_\nu) \quad (27)$$

This model is multi-linear in the odds and, in principle, agrees with the qualitative principles advocated earlier. Note that this model must also be dropped because the simulations indicate that $\psi_3 = 0$ in model M2, thus requiring $\gamma_1$ and $\gamma_2$ to be zero. Following the 3-step procedure in Section 4.4, we postulate the parallel combination

$$\gamma_1 = e_0 + \frac{e_1}{\delta_U} + \frac{e_2}{\delta_I} \text{ and } \gamma_2 = f_0 + \frac{f_1}{\delta_U} + \frac{f_2}{\delta_I} \quad (28)$$

and a similar analysis leads to

$$\gamma_1 \approx \frac{4}{\delta_U} + \frac{10}{\delta_I} \text{ and } \gamma_2 \approx \frac{12}{\delta_U} + \frac{2}{\delta_I} \quad (29)$$

The results of the fit are shown in Table 4.

The $R^2$ in Table 2 is comparable to the $R^2$ in Table 4. The final form of model M3 is

$$\rho^2 = \left(1 + \frac{2\,(2\delta_I + 5\delta_U)}{\delta_I \delta_U} O_\mu\right)\left(1 + \frac{2\,(6\delta_I + \delta_U)}{\delta_I \delta_U} O_\nu\right) \quad (30)$$

For $\nu = 0$, the model in Eq. (30) depends on $O_\mu$, and for $\mu = 0$, the model in Eq. (30) depends on $O_\nu$.

**Table 3 Regression coefficients based on regression analysis of $\psi_1$ and $\psi_2$ versus the linear densities, $\delta_U$ and $\delta_I$.**

| Regression variable | Coefficient of $(1/\delta_U)$ | Coefficient of $(1/\delta_I)$ | Intercept | $R^2$ |
|---|---|---|---|---|
| $\psi_1$ | $6.04''$ | $11.72'$ | $-0.01$ | 0.83 |
| $\psi_2$ | $12.23'$ | $5.56''$ | $-0.007$ | 0.81 |

Note: Double primes $''$ in coefficients in Table 3 indicate a standard deviation $\sigma \leqslant 0.5$ in the coefficients ($c$ and $d$) of Formula (23) as obtained from the 50 fits. A single prime $'$ indicates that $0.5 < \sigma \leqslant 0.8$. The standard deviations in the Intercept in Table 3 were less than 0.005 for all cases. All coefficients and intercept in Table 3 had very small $p$-value $p \leqslant 0.01$. The $R^{2\,[43]}$ column is used for evaluation of the regression fit.

**Table 4 Regression coefficients based on regression analysis of $\gamma_1$ and $\gamma_2$ versus the linear densities, $\delta_U$ and $\delta_I$.**

| Regression variable | Coefficient of $(1/\delta_U)$ | Coefficient of $(1/\delta_I)$ | Intercept | $R^2$ |
|---|---|---|---|---|
| $\gamma_1$ | $3.71''$ | $10.45''$ | 0.0061 | 0.85 |
| $\gamma_2$ | $12.36'$ | $2.17'$ | 0.007 | 0.81 |

Note: Double primes $''$ in Table 4 represent a standard deviation $\sigma \leqslant 0.4$ in coefficients ($e$ and $f$) of Eq. (28), as obtained from the 50 fits. Single primes $'$ indicate $0.4 < \sigma \leqslant 0.6$. The standard deviation of the Intercept in Table 4 was less than 0.005 for all cases. All coefficients and the intercept in Table 4 had very small $p$-value $p \leqslant 0.01$. The $R^{2\,[43]}$ column is used for evaluation of the regression fit.

## 6 Evaluation of the Final Models Across Different Datasets

There are models which are somewhat equivalent for very small FID and FUD, but they have different performance for large FID and FUD. A symmetrical model M1 is given by

$$\rho = 1 + \Phi\left(\frac{\delta_I + 2\delta_U}{\delta_U \delta_I} O_\mu + \frac{2\delta_I + \delta_U}{\delta_U \delta_I} O_\nu\right) \quad (31)$$

with $\Phi \approx 2.3$. Next symmetrical model M2, which is the best, is given by

$$\rho^2 = 1 + \Psi\left(\frac{\delta_I + 2\delta_U}{\delta_U \delta_I} O_\mu + \frac{2\delta_I + \delta_U}{\delta_U \delta_I} O_\nu\right) \quad (32)$$

with $\Psi \approx 5.6$. The non-symmetrical factored model M3 is given by

$$\rho^2 = \left(1 + \frac{4\delta_I + 10\delta_U}{\delta_U \delta_I} O_\mu\right) \times \left(1 + \frac{12\delta_I + 2\delta_U}{\delta_U \delta_I} O_\nu\right) \quad (33)$$

To evaluate the 3 proposed models against the original dataset, we conduct the following experiment.

(1) Construct 6 primary matrices as mentioned in Table 5. Here, $P_1$ and $P_2$ are extracted from the 1M MoviesLens dataset. $P_3$ and $P_4$ are extracted from the 25M MoviesLens dataset. $P_5$ and $P_6$ are extracted from the Yahoo! Music dataset.

**Table 5 Details of the primary rating matrices extracted for evaluating the performance of SSAL models across different datasets.**

| Primary dataset | Number of users $m$ | Number of items $n$ | Density $\delta$ | Rating scale of data | Source |
|---|---|---|---|---|---|
| $P_1$ | 6040 | 3706 | 0.045 | 1 to 5 | 1M-ML |
| $P_2$ | 4607 | 2080 | 0.095 | 1 to 5 | 1M-ML |
| $P_3$ | 8000 | 4004 | 0.101 | 0.5 to 5 | 25M-ML |
| $P_4$ | 4009 | 8017 | 0.151 | 0.5 to 5 | 25M-ML |
| $P_5$ | 3500 | 6000 | 0.08 | 1 to 100 | YM |
| $P_6$ | 5006 | 5011 | 0.12 | 1 to 100 | YM |

(2) Evaluate the models in Eqs. (31)–(33) by subsampling about 500 times from each primary matrix.

(3) Compute the MAE for either $\rho$ or $\rho^2$. The MAE of an estimate $\hat{q}$ is defined as

$$\text{MAE}\,(\hat{q}) = \frac{1}{|K|}\sum_{k\in K}|q_k - \hat{q}_k| \tag{34}$$

where $q$ is either $\rho$ or $\rho^2$, and $|K|$ is the number of subsampled matrices derived from each primary matrix.

We report the evaluation results for the models in Eq. (31)–(33) based on six different primary rating matrices as shown in Table 6. Note that the FID and the FUD were varied for a wide range from 0.1 to 0.9.

The order of performance of 3 models in Eqs. (31)–(33) to estimate SSAL is the same among the different datasets. Therefore, each model has the same performance order ($P_3$, $P_4$, $P_2$, $P_6$, $P_1$, $P_5$) for datasets while estimating ratio $\rho$ or $\rho^2$. Also, the low MAE of the AL indicator (particularly for M1 and M2) across all six datasets indicates that the proposed SSAL models are applicable to different datasets of different domains.

## 7 Validation of the Models for Different Levels of Subsampling

In this section, we will validate the 3 models developed at different levels of sub-sampling. The validations are based on the subsampled matrices of the primary matrices in Table 1. The three cases of sub-sampling levels considered are

(1) **Weak subsampling**: $\mu \leqslant 0.5, \nu \leqslant 0.5$;

(2) **Strong subsampling**: $\mu > 0.5, \nu > 0.5$; and

(3) **All subsampling**: The samples above are combined.

The validation results are summarized in Table 7. M1, M2 and M3 perform similarly for weak sampling with $R^2 \approx 0.6$ for all the models. A low $R^2$ is expected for weak subsampling because there is less variation in the SSAL for $\mu, \nu < 0.5$. Moreover, the statistical significance of the three models is justified by the low MAE($\hat{q}$) values of each model.

Table 7 also shows that the M2 model performed better than the other two models for strong subsampling ($\mu > 0.5, \nu > 0.5$) and for all subsampling. Therefore, the model M2 in Eq. (31) with $\Psi = 5.6$ will be considered superior in explaining the variations in SSAL based on the rating data characteristics.

**Table 6 MAE in estimating the ratios for the 3 models.**

| Model | SSAL indicator ($q$) | MAE of the SSAL indicator | | | | | |
|---|---|---|---|---|---|---|---|
| | $\rho$ or $\rho^2$ | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | $P_6$ |
| M1 | $\rho$ | 0.051 | 0.042 | 0.015 | 0.024 | 0.067 | 0.043 |
| M2 | $\rho^2$ | 0.110 | 0.088 | 0.026 | 0.040 | 0.13 | 0.089 |
| M3 | $\rho^2$ | 0.132 | 0.113 | 0.029 | 0.050 | 0.142 | 0.092 |

**Table 7 Evaluation of 3 models for 3 different subsampling levels.**

| Case | Model | $R^2$ | MAE ($\hat{q}$) |
|---|---|---|---|
| Weak subsampling | M1 | 0.55 | 0.018 |
| | M2 | 0.61 | 0.036 |
| | M3 | 0.59 | 0.037 |
| Strong subsampling | M1 | 0.69 | 0.061 |
| | M2 | 0.75 | 0.13 |
| | M3 | 0.46 | 0.18 |
| All subsampling | M1 | 0.71 | 0.042 |
| | M2 | 0.79 | 0.097 |
| | M3 | 0.60 | 0.12 |

The validation so far is based on $\Phi = 2.3$ in M1 in Eq. (31) and $\Psi = 5.6$ in M2 in Eq. (32). Now, we will find the best value of $\Phi$ and $\Psi$ for the three different cases of subsampling. The results shown in Table 8 indicate that the performances of the M1 and M2 are similar to the analysis as in Table 7. Also, we can see from Table 8 that $\Psi/\Phi \approx 2$ for weak FUD and FID, as expected.

Next, we randomly consider 5000 combinations of the five data characteristics $(m, n, \delta, \mu, \nu)$ and perform the analysis based on model M2 in Eq. (32) with $\Psi = 5.6$. The process was repeated for 100 times and the evaluation results of M2 is as shown in Figs. 8 and 9 which show the efficacy of model M2 to correctly estimate the SSAL based on the subsampling levels and primary matrix characteristics.

We present an example to compute $\rho^2 = \text{MSE}^S/\text{MSE}^P$ given $\mu$, $\nu$, and the linear densities of the primary rating matrix. Let us assume that 80 percent of the users and items are dropped from PRM $P$, and thus the FUD $\mu$ =0.8 and the FID $\nu = 0.8$. Then, $O_\mu$ and $O_\nu$ can be computed as

$$O_\mu = \frac{\mu}{1-\mu} = \frac{0.8}{0.2} = 4$$

**Table 8 Relation of $\Psi$ and $\Phi$ for different sub-sampling levels.**

| Case | Model | $R^2$ | MAE | Best fit | $\Psi/\Phi$ |
|---|---|---|---|---|---|
| Weak subsampling | M1 | 0.55 | 0.018 | $\Phi = 2.41$ | 2.1 |
| | M2 | 0.61 | 0.036 | $\Psi$ =5.12 | |
| Strong subsampling | M1 | 0.71 | 0.061 | $\Phi$ =2.02 | 2.99 |
| | M2 | 0.75 | 0.13 | $\Psi = 6.04$ | |
| All subsampling | M1 | 0.72 | 0.042 | $\Phi$ =1.98 | 2.88 |
| | M2 | 0.79 | 0.094 | $\Psi$ =5.72 | |

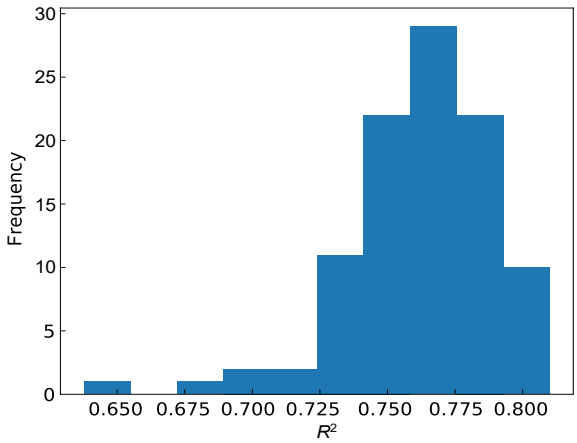


**Fig. 8 Analysis of $R^2$ of model M2 versus PMC.**

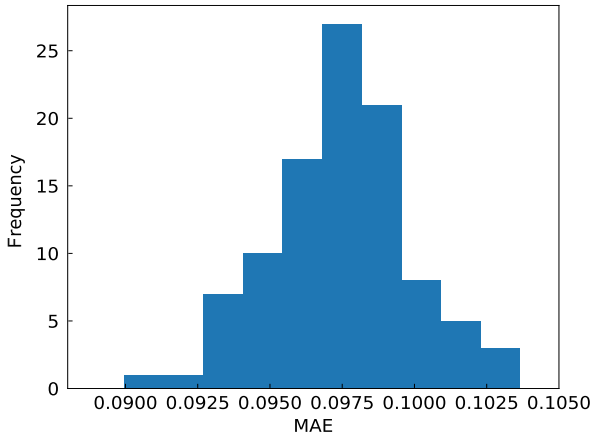


**Fig. 9 Analysis of MAE of model M2 versus PMC.**

$$O_\nu = \frac{\nu}{1-\nu} = \frac{0.8}{0.2} = 4$$

Also, we assume that, the average number of ratings per user $\delta_U$ is 20 and the average number of ratings per item $\delta_I$ is 20 in the PRM $P$. Then $\rho^2 = \text{MSE}^S/\text{MSE}^P$ can be estimated using M2 in Eq. (32) with $\Psi = 5.6$ as

$$\rho^2 = 1 + 5.6\left(\frac{20+2\times 20}{20\times 20}\times 4 + \frac{20+2\times 20}{20\times 20}\times 4\right) = 7.72$$

which means that our best model predicts that MSE of SVD CF increases by 7.7 times for any reasonable dataset.

## 8 Conclusion and Future Work

In this work, we proposed predictive SSAL models which estimate the loss in the performance of SVD CF algorithm due to subsampling. The SSAL models are constant across the datasets which suggest a theoretical underpinning. Moreover, extensive experiment showed that the SSAL models depend only on the odds ratio of dropping a user $O_\mu$, odds ratio of dropping an item $O_\nu$, and linear densities $\delta_U$ and $\delta_I$.

An SSAL model, M2 as in Eq. (32), based on MSE performed best among the proposed models. M2 is linear in $O_\mu$ and $O_\nu$ with no cross terms involved. The coefficients of $O_\mu$ and $O_\nu$ in M2 are linear in $\delta_U$ and $\delta_I$.

The SSAL models were evaluated with three well-know public datasets: (1) 1M MovieLens dataset, (2) 25M MovieLens dataset, and (3) Yahoo! Music Rating dataset. The results strongly suggest a theoretical justification or prediction for the coefficients of the SSAL model in terms of the linear densities $\delta_U$ and $\delta_I$ of the rating data. A theoretical justification would imply that the models developed here apply to other machine learning problems beyond collaborative filtering.

## References

[1] B. Smith and G. Linden, Two decades of recommender systems at Amazon.com, *IEEE Internet Comput.*, vol. 21, no. 3, pp. 12–18, 2017.

[2] C. A. Gomez-Uribe and N. Hunt, The Netflix recommender system: Algorithms, business value, and innovation, *ACM Trans. Manag. Inf. Syst.*, vol. 6, no. 4, p. 13, 2015.

[3] I. Pilászy and D. Tikk, Recommending new movies: Even a few ratings are more valuable than metadata, in *Proc. 3rd ACM Conf. on Recommender Systems*, New York, NY, USA, 2009, pp. 93–100.

[4] P. K. Singh, P. K. D. Pramanik, and P. Choudhury, Collaborative filtering in recommender systems: Technicalities, challenges, applications, and research trends, in *New Age Analytics*, G. Shrivastava, S. L. Peng, H. Bansal, K. Sharma, and M. Sharma, eds. New York, NY, USA: Apple Academic Press, 2020, pp. 183–215.

[5] J. L. Herlocker, J. A. Konstan, and J. Riedl, Explaining collaborative filtering recommendations, in *Proc. 2000 ACM Conf. on Computer Supported Cooperative Work*, Philadelphia, PA, USA, 2000, pp. 241–250.

[6] G. Adomavicius and A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

[7] Z. Liu, X. Luo, and Z. Wang, Convergence analysis of single latent factor-dependent, nonnegative, and multiplicative update-based nonnegative latent factor models, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1737–1749, 2021.

[8] D. Wu, M. Shang, X. Luo, and Z. Wang, An $L_1$-and-$L_2$-norm-oriented latent factor model for recommender systems, *IEEE Trans. Neural Netw. Learn. Syst.*, doi: 10.1109/TNNLS.2021.3071392.

[9] D. Wu, X. Luo, M. Shang, Y. He, G. Wang, and X. Wu, A data-characteristic-aware latent factor model for web services QoS prediction, *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 6, pp. 2525–2538, 2022.

[10] X. Luo, Z. Wang, and M. Shang, An instance-frequency-weighted regularization scheme for non-negative latent

factor analysis on high-dimensional and sparse data, *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 51, no. 6, pp. 3522–3532, 2021.

[11] X. Luo, W. Qin, A. Dong, K. Sedraoui, and M. Zhou, Efficient and high-quality recommendations via momentum-incorporated parallel stochastic gradient descent-based learning, *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 2, pp. 402–411, 2021.

[12] Y. Liu, T. A. N. Pham, G. Cong, and Q. Yuan, An experimental evaluation of point-of-interest recommendation in location-based social networks, *Proceedings VLDB Endowment*, vol. 10, no. 10, pp. 1010–1021, 2017.

[13] G. Adomavicius and J. Zhang, Impact of data characteristics on recommender systems performance, *ACM Trans. Manag. Inf. Syst.*, vol. 3, no. 1, p. 3, 2012.

[14] A. Bellogín and A. P. de Vries, Understanding similarity metrics in neighbour-based recommender systems, in *Proc. Conf. on the Theory of Information Retrieval*, Copenhagen, Denmark, 2013, pp. 48–55.

[15] C. Desrosiers and G. Karypis, A comprehensive survey of neighborhood-based recommendation methods, in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, eds. New York, NY, USA: Springer, 2011, pp. 107–144.

[16] F. Cacheda, V. Carneiro, D. Fernández, and V. Formoso, Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems, *ACM Trans. Web*, vol. 5, no. 1, p. 2, 2011.

[17] M. A. Ghazanfar and A. Prugel-Bennett, The advantage of careful imputation sources in sparse data-environment of recommender systems: Generating improved SVD-based recommendations, *Informatica*, vol. 37, no. 1, pp. 61–92, 2013.

[18] V. W. Anelli, T. Di Noia, E. Di Sciascio, C. Pomo, and A. Ragone, On the discriminative power of hyper-parameters in cross-validation and how to choose them, in *Proc. $13^{th}$ ACM Conf. on Recommender Systems*, Copenhagen, Denmark, 2019, pp. 447–451.

[19] E. B. Nilsen, D. E. Bowler, and J. D. C. Linnell, Exploratory and confirmatory research in the open science era, *J. Appl. Ecol.*, vol. 57, no. 4, pp. 842–847, 2020.

[20] J. Lee, M. Sun, and G. Lebanon, A comparative study of collaborative filtering algorithms, arXiv preprint arXiv: 1205.3193, 2012.

[21] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, Item-based collaborative filtering recommendation algorithms, in *Proc. 10th Int. Conf. on World Wide Web*, Hong Kong, China, 2001, pp. 285–295.

[22] V. H. Vegeborn and H. Rahmani, *Comparison and Improvement of Collaborative Filtering Algorithms*, Stockholm: KTH, 2017.

[23] Y. Deldjoo, T. Di Noia, E. Di Sciascio, and F. A. Merra, How dataset characteristics affect the robustness of collaborative recommendation models, in *Proc. $43^{rd}$ Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2020, pp. 951–960.

[24] M. Forster and E. Sober, How to tell when simpler, more unified, or less Ad hoc theories will provide more accurate predictions, *Br. J. Philos. Sci.*, vol. 45, no. 1, pp. 1–35, 1994.

[25] R. Dubin, *Theory Building*. New York, NY, USA: Free Press, 1969.

[26] M. C. Lin, A. J. T. Lee, R. T. Kao, and K. T. Chen, Stock price movement prediction using representative prototypes of financial reports, *ACM Trans. Manag. Inf. Syst.*, vol. 2, no. 3, p. 19, 2011.

[27] G. Shmueli and O. Koppius, *Predictive Analytics in Information Systems Research*, College Park: University of Maryland, 2010.

[28] S. Poudel and M. Bikdash, Optimal dependence of performance and efficiency of collaborative filtering on random stratified subsampling, *Big Data Mining and Analytics*, vol. 5, no. 3, pp. 192–205, 2022.

[29] GroupLens, MovieLens 25M dataset, https://grouplens.org/datasets/movielens/25m/, 2019.

[30] F. M. Harper and J. A. Konstan, The movielens datasets: History and context, *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, p. 19, 2016.

[31] G. H. Golub, and C. Reinsch, Singular value decomposition and least squares solutions, in *Linear Algebra*, J. H. Wilkinson and C. Reinsch, eds. Berlin, Heidelberg, Germany: Springer, 1971, pp. 134–151.

[32] N. Hug, Surprise: A python library for recommender systems, *J. Open Source Softw.*, vol. 5, no. 52, p. 2174, 2020.

[33] G. Shani and A. Gunawardana, Evaluating recommendation systems, in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira and P. B. Kantor, eds. New York, NY, USA: Springer, 2011, pp. 257–297.

[34] G. Schröder, M. Thiele, and W. Lehner, Setting goals and choosing metrics for recommender system evaluations, in *Proc. Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces*, Chicago, IL, USA, 2011, pp. 78–85.

[35] S. Poudel, A study of disease diagnosis using machine learning, presented at the 2nd Int. Electronic Conf. on Healthcare, Basel, Switzerland, 2022.

[36] M. Jalili, S. Ahmadian, M. Izadi, P. Moradi, and M. Salehi, Evaluating collaborative filtering recommender algorithms: A survey, *IEEE Access*, vol. 6, pp. 74003–74024, 2018.

[37] S. Poudel, Improving collaborative filtering recommendation systems via optimal sub-sampling and aspect-based interpretability, PhD dissertation, North Carolina Agricultural and Technical State University, Greensboro, NC, USA, 2022.

[38] T. Chai and R. R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature, *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014.

[39] J. Frost, Mean squared error (MSE), https://statisticsbyjim.com/regression/mean-squared-error-mse/, 2022.

[40] GroupLens, MovieLens 1M dataset, https://grouplens.org/datasets/movielens/1m/, 2003.
[41] Webscope | Yahoo labs, https://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=1, 2019.
[42] E. C. Alexopoulos, Introduction to multivariate regression analysis, *Hippokratia*, vol. 14, no. Suppl 1, pp. 23–28, 2010.
[43] K. Kumari and S. Yadav, Linear regression analysis study, *J. Pract. Cardiovasc. Sci.*, vol. 4, no. 1, pp. 33–36, 2018.

**Marwan Bikdash** received the MS and PhD degrees in electrical engineering from Virginia Tech, USA in 1990 and 1993, respectively. He is currently a professor and the chair of the Department of Computational Data Science and Engineering, North Carolina A&T State University. He teaches and conducts research in signals and systems, computational intelligence, and modeling and simulations of systems with applications in health, energy, and engineering. He has authored over 130 journal and conference papers. He has supported, advised, and graduated over 50 MS and PhD students. His projects have been funded by the Jet Propulsion Laboratory, Defense Threat Reduction Agency, Army Research Lab, NASA, National Science Foundation, the Office of Naval Research, Boeing Inc., Hewlett Packard, National Renewable Energy Laboratories, the Army Construction Engineering Research Laboratory, etc.

**Samin Poudel** received the MS degree in physics and the PhD degree in computational data science and engineering from North Carolina A&T State University, USA in 2017 and 2022, respectively. His research interests include but not limited to data analytics, data mining, machine learning, developing models, and optimizing techniques based on data.