Low-Rank Matrix Approximation with Stability

Dongsheng Li[§] LDSLI@CN.IBM.COM
Chao Chen[†] CHENCH.RESCH@GMAIL.COM
Qin Lv[‡] QIN.LV@COLORADO.EDU
Junchi Yan[§] YANJC@CN.IBM.COM
Li Shang[‡] LI.SHANG@COLORADO.EDU
Stephen M. Chu[§] SCHU@US.IBM.COM
[§]IBM Research - China, 399 Keyuan Road, Shanghai P. R. China 201203
[†]Tongji University, 4800 Caoan Road, Shanghai P.R. China 201804
[‡]University of Colorado Boulder, Boulder, Colorado USA 80309

Abstract

Low-rank matrix approximation has been widely adopted in machine learning applications with sparse data, such as recommender systems. However, the sparsity of the data, incomplete and noisy, introduces challenges to the algorithm stability – small changes in the training data may significantly change the models. As a result, existing low-rank matrix approximation solutions yield low generalization performance, exhibiting high error variance on the training dataset, and minimizing the training error may not guarantee error reduction on the testing dataset. In this paper, we investigate the algorithm stability problem of low-rank matrix approximations. We present a new algorithm design framework, which (1) introduces new optimization objectives to guide stable matrix approximation algorithm design, and (2) solves the optimization problem to obtain stable low-rank approximation solutions with good generalization performance. Experimental results on real-world datasets demonstrate that the proposed work can achieve better prediction accuracy compared with both state-of-the-art low-rank matrix approximation methods and ensemble methods in recommendation task.

1. Introduction

Low-rank matrix approximation (LRMA) has been widely adopted in machine learning applications with sparse data, e.g., recommender systems (Paterek, 2007; Koren et al., 2009). In LRMA-based recommender systems (Paterek, 2007; Koren et al., 2009), a given user-item rating matrix is approximated using observed ratings (often sparse), then user ratings on unrated items are predicted using the dot product of corresponding user and item feature vectors. LRMA methods have the capability of reducing the dimensionality of user/item rating vectors, hence are suitable to handle applications with sparse data (Koren et al., 2009). Indeed, LRMA-based solutions have been widely used in existing recommender systems (Paterek, 2007; Koren et al., 2009; Lee et al., 2013; Beutel et al., 2015).

The sparsity of the data, incomplete and noisy (Keshavan et al., 2010; Candès & Recht, 2012), introduces challenges to the algorithm stability. In LRMA, models are biased to the limited training data (sparse), so that small changes in the training data (noisy) may significantly change the models. As demonstrated in this work, existing LRMA methods cannot provide stable matrix approximations. Such unstable matrix approximations will introduce high training error variance, and minimizing the training error may not guarantee consistent error reduction on the testing dataset, i.e., low generalization performance (Bousquet & Elisseeff, 2001; Srebro et al., 2004a;b). In other words, the algorithm stability has direct impact on generalization performance, and an unstable LRMA algorithm has low generalization performance (Srebro et al., 2004a).

Heuristic techniques, such as cross-validation and ensemble learning (Koren, 2008; Mackey et al., 2011; Lee et al., 2013; Chen et al., 2015), can be adopted to improve the generalization performance of LRMA. However, cross validation methods have the drawback that the amount of data available for model learning is reduced (Kohavi, 1995; Bousquet & Elisseeff, 2001). Ensemble LRMA methods (Lee et al., 2013; Chen et al., 2015) are computationally expensive due to the training of sub-models. Recently,

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

the notion of “algorithmic stability” has been introduced to investigate the theoretical bound of the generalization performance of learning algorithms (Bousquet & Elisseeff, 2001; 2002; Agarwal & Niyogi, 2009; Shalev-Shwartz et al., 2010; London et al., 2013). It is timely to develop stable algorithms with low generation errors, suitable for learning applications with sparse data.

This paper presents a stable LRMA algorithm design framework. It formulates new optimization objectives to derive stable LRMA algorithms, namely SMA, and solves the new optimization objectives to obtain SMA solutions with good generalization performance. We first introduce the stability notion in LRMA, and then develop theoretical guidelines for deriving LRMA solutions with high stability. Then, we formulate a new optimization problem for achieving stable LRMA, in which minimizing the loss function can obtain solutions with high stability, i.e., good generalization performance. Finally, we develop a stochastic gradient descent method to solve the new optimization problem. Experimental results on real-world datasets demonstrate that the proposed SMA method can deliver a stable LRMA algorithm, which achieves better prediction accuracy over state-of-the-art single LRMA methods and ensemble LRMA methods in recommendation task. The key contributions of this paper are as follows: (1) this work first introduces the stability concept in LRMA, which can provide theoretical guidelines for deriving stable matrix approximation; (2) a stable LRMA algorithm design framework is proposed, which can achieve high stability, i.e., high generalization performance by designing and solving new optimization objectives derived based on stability analysis; (3) evaluation using real-world datasets demonstrates that the proposed work can make significant improvement in prediction accuracy over state-of-the-art LRMA methods and ensemble methods in recommendation task.

The rest of this paper is organized as follows: Section 2 formulates stability problem in LRMA and formally proves key observations. Section 3 presents details of SMA. Section 4 presents the experimental results. Section 5 discusses related work, and we conclude this work in Section 6.

2. Stability of LRMA

This section first summarizes LRMA, and then introduces the definition of stability. Next, we conduct quantitative analysis of the relationship between algorithm stability and generalization error. Finally, we present key guidelines to developing stable LRMA, and derive theoretical proof.

2.1. Low-Rank Matrix Approximation

In this paper, upper case letters, such as R, U, V denote matrices. For a targeted matrix $R \in \mathbb{R}^{m \times n}$, Ω denotes the set of observed entries in R, and $\hat{R}$ denotes the low-rank approximation of R. The objective of r-rank matrix approximation is to determine two feature matrices, i.e., $U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}$, s.t., $\hat{R} = UV^T$. The rank r is considered low in many scenarios, because $r \ll \min\{m, n\}$ can deliver good performance in many real applications.

The feature matrices U and V cannot be determined arbitrarily in low-rank matrix approximation. Generally, loss functions should be defined towards different tasks, and U and V are chosen to minimize such loss functions (Lee & Seung, 2001; Salakhutdinov & Mnih, 2007; 2008; Yan et al., 2010). Let loss function $Loss(R, \hat{R})$ be the error of approximating R by $\hat{R}$, the optimization problem of LRMA can be formally described as follows:

$$\hat{R} = \arg\min_{X} Loss(R, X), \ rank(X) = r. \tag{1}$$

The loss functions should vary for different tasks. For instance, Singular Value Decomposition (SVD) usually adopts Frobenius norm to define loss function, and Compressed Sensing adopts nuclear norm. Typically, the problems defined by Equation 1 are often difficult non-convex optimization problems, so that iterative methods, e.g., gradient descent (GD), stochastic gradient descent (SGD), are adopted to find solutions that will converge to local minimum (Lee & Seung, 2001; Salakhutdinov & Mnih, 2007).

2.2. Stability w.r.t Matrix Approximation

Recent work on algorithmic stability (Bousquet & Elisseeff, 2001; 2002; Lan et al., 2008; London et al., 2013) demonstrated that a stable learning algorithm has the property that replacing one element in the training set does not result in significant change to the algorithm's output. Therefore, if we take the training error as a random variable, the training error of stable learning algorithm should have small variance. This implies that stable algorithms have the property that the training errors are close to the test errors (Bousquet & Elisseeff, 2001; Lan et al., 2008; London et al., 2013). The rest of this section introduces and analyzes the algorithm stability problem of low-rank matrix approximation.

Root Mean Square Error (RMSE), as a common evaluation metric for recommendation tasks, is adopted to measure the stability of matrix approximation. Let $\mathcal{D}(\hat{R}) = \sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(R_{i,j} - \hat{R}_{i,j})^2}$ be the root mean square error of approximating R with $\hat{R}$. One of the objectives of matrix approximation is to approximate a given matrix R based on a set of observed entries Ω ($\mathcal{D}_\Omega(\hat{R}) = \sqrt{\frac{1}{|\Omega|}\sum_{(i,j)\in\Omega}(R_{i,j} - \hat{R}_{i,j})^2}$). Thus, the stability of approximating R is defined as follows.

Definition 1 (Stability w.r.t. Matrix Approximation)**.** *For any $R \in \mathbb{F}^{m \times n}$, choose a subset of entries Ω from R uniformly. For a given $\epsilon > 0$, we say that $\mathcal{D}_\Omega(\hat{R})$ is δ-stable if* *the following holds:*

$$\Pr[|\mathcal{D}(\hat{R}) - \mathcal{D}_{\Omega}(\hat{R})| \leq \epsilon] \geq 1 - \delta. \quad (2)$$

Matrix approximation with stability guarantee has the property that the generalization error is bounded. Minimizing the training error will have a high probability of minimizing the test error. The stability notion introduced in this work defines how stable an approximation is in terms of the overall prediction error. It is different from the Uniform Stability definition (Bousquet & Elisseeff, 2001), which defines the prediction stability on individual entries. This new stability notion makes it possible to measure the generalization performance between different approximations. For instance, for any two subsets of entries Ω_1 and Ω_2 from R, approximating R by Ω_1 and Ω_2 are δ_1-stable and δ_2-stable, respectively, then $\mathcal{D}_{\Omega_1}(\hat{R})$ is more stable than $\mathcal{D}_{\Omega_2}(\hat{R})$ if $\delta_1 < \delta_2$. This implies that $\mathcal{D}_{\Omega_1}(\hat{R})$ is close to $\mathcal{D}(\hat{R})$ with higher probability than $\mathcal{D}_{\Omega_2}(\hat{R})$, i.e., minimizing $\mathcal{D}_{\Omega_1}(\hat{R})$ will lead to solutions that are of higher probabilities with better generalization performance than minimizing $\mathcal{D}_{\Omega_2}(\hat{R})$. In summary, using the stability notion introduced in this paper, we can define new matrix approximation problems which can yield solutions with high stability, i.e., high generalization performance.

2.3. Stability vs. Generalization Error

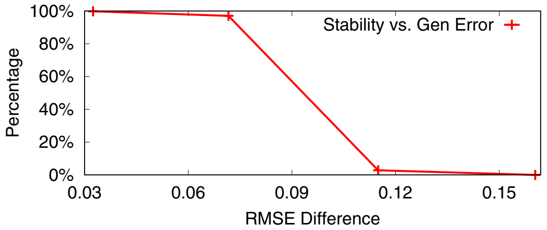

Figure 1. Stability vs. generalization error with different rank r on MovieLens (1M) dataset.

Figure 1 quantifies stability changes of LRMA method with the generalization error when rank r increases from 5 to 20. This experiment uses RSVD (Paterek, 2007), a popular LRMA-based recommendation algorithm, on the MovieLens (1M) dataset ($\sim 10^6$ ratings, 6,040 users, 3,706 items). We choose ϵ in Definition 1 as 0.0046 to cover all error differences when $r = 5$. We compute $\Pr[|\mathcal{D}(\hat{R}) - \mathcal{D}_{\Omega}(\hat{R})| \leq \epsilon]$ with 500 different runs to measure stability (y-axis), and compute RMSE differences between training and test data to measure generalization error (x-axis).

As shown in Figure 1, the generalization error increases when rank r increases, because LRMA models become more complex and more biased on training data with higher ranks. In contrary, the stability of RSVD decreases with r increases. This indicates that (1) stability decreases with generalization error increases, and (2) RSVD cannot provide stable recommendation even the rank is as low as 20. This study demonstrates that existing LRMA methods suffer from low generalization performance due to low algorithm stability. Therefore, it is important to develop stable LRMA methods with good generalization performance.

2.4. Stability Analysis

Next, we introduce the Hoeffding's Lemma, and then analyze the the Stability properties of low-rank matrix approximation problems.

Lemma 1 (Hoeffding's Lemma)**.** *Let X be a real-valued random variable with zero mean and $\Pr(X \in [a, b]) = 1$. Then, for any $s \in \mathbb{R}$, $\mathrm{E}\left[e^{sX}\right] \leq \exp\left(\frac{1}{8}s^2(b-a)^2\right)$.*

Following the Uniform Stability (Bousquet & Elisseeff, 2001), given a stable LRMA algorithm, the approximation results remain stable if the change of the training data set, i.e., the set of observed entries Ω from the original matrix R, is small. For instance, we can remove a subset of easily predictable entries from Ω to obtain Ω'. It is desirable that the solution of minimizing both $\mathcal{D}_{\Omega}$ and $\mathcal{D}_{\Omega'}$ together will be more stable than the solution of minimizing $\mathcal{D}_{\Omega}$ only. The following Theorem 1 formally proves the statement.

Theorem 1. *Let Ω ($|\Omega| > 2$) be a set of observed entries in R. Let $\omega \subset \Omega$ be a subset of observed entries, which satisfy that $\forall (i,j) \in \omega$, $|R_{i,j} - \hat{R}_{i,j}| \leq \mathcal{D}_{\Omega}(\hat{R})$. Let $\Omega' = \Omega - \omega$, then for any $\epsilon > 0$ and $1 > \lambda_0, \lambda_1 > 0$ ($\lambda_0 + \lambda_1 = 1$), $\lambda_0\mathcal{D}_{\Omega}(\hat{R}) + \lambda_1\mathcal{D}_{\Omega'}(\hat{R})$ and $\mathcal{D}_{\Omega}(\hat{R})$ are δ_1-stable and δ_2-stable, resp., then $\delta_1 \leq \delta_2$.*

Proof. Let's assume that $\mathcal{D}(\hat{R}) - \mathcal{D}_{\Omega}(\hat{R}) \in [-a, a]$ ($a = \sup\{\mathcal{D}(\hat{R}) - \mathcal{D}_{\Omega}(\hat{R})\}$) and $\mathcal{D}(\hat{R}) - (\lambda_0\mathcal{D}_{\Omega}(\hat{R}) + \lambda_1\mathcal{D}_{\Omega'}(\hat{R})) \in [-a', a']$ ($a' = \sup\{\mathcal{D}(\hat{R}) - (\lambda_0\mathcal{D}_{\Omega}(\hat{R}) + \lambda_1\mathcal{D}_{\Omega'}(\hat{R}))\}$) are two random variables with 0 mean.

Based on Markov's inequality, for any $t > 0$, we have

$$\Pr[\mathcal{D}(\hat{R}) - \mathcal{D}_{\Omega}(\hat{R}) \geq \epsilon] \leq \frac{E(e^{t(\mathcal{D}(\hat{R}) - \mathcal{D}_{\Omega}(\hat{R}))})}{e^{t\epsilon}}.$$

Then, based on Lemma 1, we have $\Pr[\mathcal{D}(\hat{R}) - \mathcal{D}_{\Omega}(\hat{R}) \geq \epsilon] \leq \frac{\exp(\frac{1}{2}t^2a^2)}{\exp(t\epsilon)}$ and $\Pr[-\mathcal{D}(\hat{R}) + \mathcal{D}_{\Omega}(\hat{R}) \geq \epsilon] \leq \frac{\exp(\frac{1}{2}t^2a^2)}{\exp(t\epsilon)}$. Combining above two equations, we have $\Pr[|\mathcal{D}(\hat{R}) - \mathcal{D}_{\Omega}(\hat{R})| \geq \epsilon] \leq \frac{2\exp(\frac{1}{2}t^2a^2)}{\exp(t\epsilon)}$, i.e., $\Pr[|\mathcal{D}(\hat{R}) - \mathcal{D}_{\Omega}(\hat{R})| \leq \epsilon] \geq 1 - \frac{2\exp(\frac{1}{2}t^2a^2)}{\exp(t\epsilon)}$. Similarly, we have $\Pr[|\mathcal{D}(\hat{R}) - (\lambda_0\mathcal{D}_{\Omega}(\hat{R}) + \lambda_1\mathcal{D}_{\Omega'}(\hat{R}))| \leq \epsilon] \geq 1 - \frac{2\exp(\frac{1}{2}t^2a'^2)}{\exp(t\epsilon)}$. We can compare a' with a as follows:

$$\begin{aligned} a' &= \sup\{\mathcal{D}(\hat{R}) - \mathcal{D}_{\Omega}(\hat{R}) + \lambda_1(\mathcal{D}_{\Omega}(\hat{R}) - \mathcal{D}_{\Omega'}(\hat{R}))\} \\ &= \sup\{\mathcal{D}(\hat{R}) - \mathcal{D}_{\Omega}(\hat{R})\} + \lambda_1 \sup\{\mathcal{D}_{\Omega}(\hat{R}) - \mathcal{D}_{\Omega'}(\hat{R})\} \\ &= a + \lambda_1 \sup\{\mathcal{D}_{\Omega}(\hat{R}) - \mathcal{D}_{\Omega'}(\hat{R})\}. \end{aligned}$$

Since $\forall (i,j) \in \omega$, $|R_{i,j} - \hat{R}_{i,j}| \leq \mathcal{D}_\Omega(\hat{R})$, we have $1/|\omega| \sum_{(i,j)\in\omega} (R_{i,j} - \hat{R}_{i,j})^2 \leq \mathcal{D}^2_\Omega(\hat{R})$, i.e., $\mathcal{D}_\omega(\hat{R}) \leq \mathcal{D}_\Omega(\hat{R})$. Then, since $\Omega = \omega \cup \Omega'$, we have $\mathcal{D}_{\Omega'}(\hat{R}) \geq \mathcal{D}_\Omega(\hat{R})$. This means that $\sup\{\mathcal{D}_\Omega(R) - \mathcal{D}_{\Omega'}(R)\} \leq 0$. Thus, we have $a' \leq a$. This turns out that $\frac{2\exp(\frac{1}{2}t^2a'^2)}{\exp(t\epsilon)} \leq \frac{2\exp(\frac{1}{2}t^2a^2)}{\exp(t\epsilon)}$, i.e., $\delta_1 \leq \delta_2$. □

Remark. The above Theorem 1 indicates that, if we remove a subset of entries that are easier to predict than average from Ω to form Ω', then $\lambda_0\mathcal{D}_\Omega(\hat{R}) + \lambda_1\mathcal{D}_{\Omega'}(\hat{R})$ has a higher probability of being close to $\mathcal{D}(\hat{R})$ than $\mathcal{D}_\Omega(\hat{R})$. Therefore, minimizing $\lambda_0\mathcal{D}_\Omega(\hat{R}) + \lambda_1\mathcal{D}_{\Omega'}(\hat{R})$ will lead to solutions that have better generalization performance than minimizing $\mathcal{D}_\Omega(\hat{R})$. It should be noted that the condition that $\forall (i,j) \in \omega$, $|R_{i,j} - \hat{R}_{i,j}| \leq \mathcal{D}_\Omega(\hat{R})$ is not necessary. The conclusion will be the same if $\mathcal{D}_\omega(\hat{R}) \leq \mathcal{D}_\Omega(\hat{R})$. The following Proposition 1 formally shows this.

Proposition 1. *Let Ω ($|\Omega| > 2$) be a set of observed entries in R. Let $\omega \subset \Omega$ be a subset of observed entries, which satisfy that $D_\omega(\hat{R}) \leq D_\Omega(\hat{R})$. Let $\Omega' = \Omega - \omega$, then for any $\epsilon > 0$ and $1 > \lambda_0, \lambda_1 > 0$ ($\lambda_0 + \lambda_1 = 1$), $\lambda_0\mathcal{D}_\Omega(\hat{R}) + \lambda_1\mathcal{D}_{\Omega'}(\hat{R})$ and $\mathcal{D}_\Omega(\hat{R})$ are δ_1-stable and δ_2-stable, resp., then $\delta_1 \leq \delta_2$.*

Proof. This proof is omitted as it is similar to that of Theorem 1. □

However, Theorem 1 and Proposition 1 only prove that it is beneficial to remove easily predictable entries from Ω to obtain Ω', but does not show how many entries we should remove from Ω. The following Theorem 2 shows that removing more entries that satisfy $|R_{i,j} - \hat{R}_{i,j}| \leq \mathcal{D}_\Omega(\hat{R})$ can yield better Ω'.

Theorem 2. *Let Ω ($|\Omega| > 2$) be a set of observed entries in R. Let $\omega_2 \subset \omega_1 \subset \Omega$, and ω_1 and ω_2 satisfy that $\forall (i,j) \in \omega_1(\omega_2)$, $|R_{i,j} - \hat{R}_{i,j}| \leq \mathcal{D}_\Omega(\hat{R})$. Let $\Omega_1 = \Omega - \omega_1$ and $\Omega_2 = \Omega - \omega_2$, then for any $\epsilon > 0$ and $1 > \lambda_0, \lambda_1 > 0$ ($\lambda_0 + \lambda_1 = 1$), $\lambda_0\mathcal{D}_\Omega(\hat{R}) + \lambda_1\mathcal{D}_{\Omega_1}(\hat{R})$ and $\lambda_0\mathcal{D}_\Omega(\hat{R}) + \lambda_1\mathcal{D}_{\Omega_2}(\hat{R})$ are δ_1-stable and δ_2-stable, resp., then $\delta_1 \leq \delta_2$.*

Proof. Similar to Theorem 1, let's assume that $\mathcal{D}(\hat{R}) - (\lambda_0\mathcal{D}_\Omega(\hat{R}) + \lambda_1\mathcal{D}_{\Omega_1}(\hat{R})) \in [-a_1, a_1]$ ($a_1 = \sup\{\mathcal{D}(\hat{R}) - (\lambda_0\mathcal{D}_\Omega(\hat{R}) + \lambda_1\mathcal{D}_{\Omega_1}(\hat{R}))\}$) and $\mathcal{D}(\hat{R}) - (\lambda_0\mathcal{D}_\Omega(\hat{R}) + \lambda_1\mathcal{D}_{\Omega_2}(\hat{R})) \in [-a_2, a_2]$ ($a_2 = \sup\{\mathcal{D}(\hat{R}) - (\lambda_0\mathcal{D}_\Omega(\hat{R}) + \lambda_1\mathcal{D}_{\Omega_2}(\hat{R}))\}$) are two random variables with 0 mean.

Applying Lemma 1 and the Markov's inequality, we have $\Pr[|\mathcal{D}(\hat{R}) - (\lambda_0\mathcal{D}_\Omega(\hat{R}) + \lambda_1\mathcal{D}_{\Omega_1}(\hat{R}))| \leq \epsilon] \geq 1 - \frac{2\exp(\frac{1}{2}t^2a_1^2)}{\exp(t\epsilon)}$ and $\Pr[|\mathcal{D}(\hat{R}) - (\lambda_0\mathcal{D}_\Omega(\hat{R}) + \lambda_1\mathcal{D}_{\Omega_2}(\hat{R}))| \leq \epsilon] \geq 1 - \frac{2\exp(\frac{1}{2}t^2a_2^2)}{\exp(t\epsilon)}$. Since $\forall (i,j) \in \omega_1(\omega_2)$, $|R_{i,j} - \hat{R}_{i,j}| \leq \mathcal{D}_\Omega(\hat{R})$ and $\omega_2 \subset \omega_1$, we have $\mathcal{D}_{\Omega_1}(\hat{R}) \geq \mathcal{D}_{\Omega_2}(\hat{R})$. Thus, we have $\sup\{\mathcal{D}_{\Omega_1}(\hat{R}) - \mathcal{D}_{\Omega_2}(\hat{R})\} \leq 0$. Since $a_1 = a_2 + \lambda_1 \sup\{\mathcal{D}_{\Omega_1}(\hat{R}) - \mathcal{D}_{\Omega_2}(\hat{R})\}$, we have $a_1 \leq a_2$. Then, similar to Theorem 1, we can conclude that $\delta_1 \leq \delta_2$. □

Remark. From Theorem 2, we know that removing more entries that are easy to predict will yield more stable matrix approximation. Therefore, it is desirable to choose Ω' as the whole set of entries which are harder to predict than average, i.e., the whole set of entries satisfying $\forall (i,j) \in \Omega'$, $|R_{i,j} - \hat{R}_{i,j}| \geq \mathcal{D}_\Omega(\hat{R})$.

Theorem 1 and 2 only consider choosing one Ω' to find stable matrix approximations. The next question is, is it possible to choose more than one Ω' that satisfy the above condition, and yield more stable solutions by minimizing them all together. The following Theorem 3 shows that incorporating K such entry sets (Ω') will be more stable than incorporating any $K-1$ out of the K entry sets.

Theorem 3. *Let Ω ($|\Omega| > 2$) be a set of observed entries in R. $\omega_1, ..., \omega_K \subset \Omega$ ($K > 1$) satisfy that $\forall (i,j) \in \omega_k$ ($1 \leq k \leq K$), $|R_{i,j} - \hat{R}_{i,j}| \leq \mathcal{D}_\Omega(\hat{R})$. Let $\Omega_k = \Omega - \omega_k$ for all $1 \leq k \leq K$. Then, for any $\epsilon > 0$ and $1 > \lambda_0, \lambda_1, ..., \lambda_K > 0$ ($\sum_{i=0}^K \lambda_i = 1$), $\lambda_0\mathcal{D}_\Omega(\hat{R}) + \sum_{k\in[1,K]} \lambda_k\mathcal{D}_{\Omega_k}(\hat{R})$ and $(\lambda_0 + \lambda_K)\mathcal{D}_\Omega(\hat{R}) + \sum_{k\in[1,K-1]} \lambda_k\mathcal{D}_{\Omega_k}(\hat{R})$ are δ_1-stable and δ_2-stable, resp., then $\delta_1 \leq \delta_2$.*

Proof. Let's assume that $\mathcal{D}(\hat{R}) - ((\lambda_0 + \lambda_K)\mathcal{D}_\Omega(\hat{R}) + \sum_{k\in[1,K-1]} \lambda_k\mathcal{D}_{\Omega_k}(\hat{R})) \in [-a, a]$ ($a = \sup\{\mathcal{D}(\hat{R}) - ((\lambda_0+\lambda_K)\mathcal{D}_\Omega(\hat{R}) + \sum_{k\in[1,K-1]} \lambda_k\mathcal{D}_{\Omega_k}(\hat{R}))\}$ and $\mathcal{D}(\hat{R}) - (\lambda_0\mathcal{D}_\Omega(\hat{R}) + \sum_{k\in[1,K]} \lambda_k\mathcal{D}_{\Omega_k}(\hat{R})) \in [-a', a']$ ($a' = \sup\{\mathcal{D}(\hat{R}) - (\lambda_0\mathcal{D}_\Omega(\hat{R}) + \sum_{k\in[1,K]} \lambda_k\mathcal{D}_{\Omega_k}(\hat{R}))\}$ are two random variables with 0 mean.

Applying Lemma 1 and the Markov's inequality, we have $\Pr[|\mathcal{D}(\hat{R}) - (\lambda_0\mathcal{D}_\Omega(\hat{R}) + \sum_{k\in[1,K]} \lambda_k\mathcal{D}_{\Omega_k}(\hat{R}))| \leq \epsilon] \geq 1 - \frac{2\exp(\frac{1}{2}t^2a'^2)}{\exp(t\epsilon)}$ and $\Pr[|\mathcal{D}(\hat{R}) - ((\lambda_0 + \lambda_K)\mathcal{D}_\Omega(\hat{R}) + \sum_{k\in[1,K-1]} \lambda_k\mathcal{D}_{\Omega_k}(\hat{R}))| \leq \epsilon] \geq 1 - \frac{2\exp(\frac{1}{2}t^2a^2)}{\exp(t\epsilon)}$. Similar as the above proofs, we have $\mathcal{D}_{\Omega_K}(\hat{R}) \leq \mathcal{D}_\Omega(\hat{R})$, which means $\sup\{\mathcal{D}_\Omega(\hat{R}) - \mathcal{D}_{\Omega_K}(\hat{R})\} \leq 0$. Since $a' = a + \lambda_K \sup\{\mathcal{D}_\Omega(\hat{R}) - \mathcal{D}_{\Omega_K}(\hat{R})\}$, we know that $a' \leq a$. Thus, we can conclude that $\frac{2\exp(\frac{1}{2}t^2a'^2)}{\exp(t\epsilon)} \leq \frac{2\exp(\frac{1}{2}t^2a^2)}{\exp(t\epsilon)}$, i.e., $\delta_1 \leq \delta_2$. □

Remark. Theorem 3 shows that minimizing $\mathcal{D}_\Omega$ together with the RMSEs of more than one hard predictable subsets of Ω will help generate more stable matrix approximation solutions.

However, sometimes it is expensive to find such "hard predictable subsets", because we do not know which subset of entries to choose without any prior knowledge. Thus, we propose a solution to obtain hard predictable subsets of Ω based on only one set of easily predictable entries as follows: 1) choose $\omega \subset \Omega$, which satisfies that $\forall (i,j) \in \omega$, $|R_{i,j} - \hat{R}_{i,j}| \leq \mathcal{D}_{\Omega}(\hat{R})$, and choose $\Omega_0 = \Omega - \omega$; 2) divide ω into K non-overlapping subsets $\omega_1, ..., \omega_K$ with the condition that $\cup_{k \in [1,K]} \omega_k = \omega$, and choose $\Omega_k = \Omega - \omega_k$ for all $1 \leq k \leq K$; and 3) minimize $\lambda_0 \mathcal{D}_{\Omega}(\hat{R}) + \sum_{k=1}^{K} \lambda_k \mathcal{D}_{\Omega_k}(\hat{R})$ to find stable matrix approximation solutions. The following Theorem 4 proves that it is desirable to minimize $\lambda_0 \mathcal{D}_{\Omega}(\hat{R}) + \sum_{k=1}^{K} \lambda_k \mathcal{D}_{\Omega_k}(\hat{R})$ instead of $\lambda_0 \mathcal{D}_{\Omega}(\hat{R}) + (1 - \lambda_0) \mathcal{D}_{\Omega_0}(\hat{R})$.

Theorem 4. *Let Ω ($|\Omega| > 2$) be a set of observed entries in R. Choose $\omega \subset \Omega$, which satisfies that $\forall (i,j) \in \omega$, $|R_{i,j} - \hat{R}_{i,j}| \leq \mathcal{D}_{\Omega}(\hat{R})$. And divide ω into K non-overlapping subsets $\omega_1, ..., \omega_K$ with the condition that $\cup_{k \in [1,K]} \omega_k = \omega$. Let $\Omega_0 = \Omega - \omega$ and $\Omega_k = \Omega - \omega_k$ for all $1 \leq k \leq K$. Then, for any $\epsilon > 0$ and $1 > \lambda_0, \lambda_1, ..., \lambda_K > 0$ ($\sum_{i=0}^{K} \lambda_i = 1$), $\lambda_0 \mathcal{D}_{\Omega}(\hat{R}) + \sum_{k=1}^{K} \lambda_k \mathcal{D}_{\Omega_k}(\hat{R})$ and $\lambda_0 \mathcal{D}_{\Omega}(\hat{R}) + (1 - \lambda_0) \mathcal{D}_{\Omega_0}(\hat{R})$ are δ_1-stable and δ_2-stable, resp., then $\delta_1 \leq \delta_2$.*

Proof. Let's first assume that $\mathcal{D}(\hat{R}) - (\lambda_0 \mathcal{D}_{\Omega}(\hat{R}) + \sum_{k=1}^{K} \lambda_k \mathcal{D}_{\Omega_k}(\hat{R})) \in [-a_1, a_1]$ ($a_1 = \sup\{\mathcal{D}(\hat{R}) - (\lambda_0 \mathcal{D}_{\Omega}(\hat{R}) + \sum_{k=1}^{K} \lambda_k \mathcal{D}_{\Omega_k}(\hat{R}))\}$ and $\mathcal{D}(\hat{R}) - (\lambda_0 \mathcal{D}_{\Omega}(\hat{R}) + (1 - \lambda_0) \mathcal{D}_{\Omega_0}(\hat{R})) \in [-a_2, a_2]$ ($a_2 = \sup\{\mathcal{D}(\hat{R}) - ((1 - \lambda_0) \mathcal{D}_{\Omega_0}(\hat{R}))\}$ are two random variables with 0 mean.

We have $\Pr[|\mathcal{D}(\hat{R}) - (\lambda_0 \mathcal{D}_{\Omega}(\hat{R}) + \sum_{k=1}^{K} \lambda_k \mathcal{D}_{\Omega_k}(\hat{R}))| \leq \epsilon] \geq 1 - \frac{2 \exp(\frac{1}{2} t^2 a_1^2)}{\exp(t\epsilon)}$ and $\Pr[|\mathcal{D}(\hat{R}) - ((1 - \lambda_0) \mathcal{D}_{\Omega_0}(\hat{R}))| \leq \epsilon] \geq 1 - \frac{2 \exp(\frac{1}{2} t^2 a_2^2)}{\exp(t\epsilon)}$. $\forall k \in [1, K]$ $\omega_k \subset \omega$ and $\forall (i,j) \in \omega$, $|R_{i,j} - \hat{R}_{i,j}| \leq \mathcal{D}_{\Omega}(\hat{R})$, we have for all $k \in [1, K]$, $\mathcal{D}_{\Omega_k} \leq \mathcal{D}_{\Omega_0}$. Sum the above inequation over all $k \in [1, K]$, we have $\sum_{k=1}^{K} \lambda_k \mathcal{D}_{\Omega_k} \leq \sum_{k=1}^{K} \lambda_k \mathcal{D}_{\Omega_0} = (1 - \lambda_0) \mathcal{D}_{\Omega_0}$. Thus, $\sup\{\mathcal{D}(\hat{R}) - (\lambda_0 \mathcal{D}_{\Omega}(\hat{R}) + \sum_{k=1}^{K} \lambda_k \mathcal{D}_{\Omega_k}(\hat{R}))\} \leq \sup\{\mathcal{D}(\hat{R}) - ((1 - \lambda_0) \mathcal{D}_{\Omega_0}(\hat{R}))\}$, i.e., $a_1 \leq a_2$. Thus, we can conclude that $\delta_1 \leq \delta_2$. □

Remark. Theorem 4 shows that if we can find only one subset of entries that are easier to predict than average, then we can probe this subset of entries to increase the stability of matrix approximations.

3. SMA Algorithm

This section presents the algorithm stability optimization problem for matrix approximation. Then, we propose a method to find out entry sets that are harder to predict than average, which is a key step for constructing this optimization problem. Finally, we present how to solve the algorithm stability optimization problem using a stochastic gradient descent method.

3.1. Model Formulation

Singular value decomposition (SVD) is one of the commonly used methods for low-rank matrix approximation (Candès & Plan, 2010). Based on the analysis of stable matrix approximation described in the previous section, it is desirable to minimize the loss functions that will lead to solutions with good generalization performance. Let $\{\Omega_1, ..., \Omega_K\}$ be subsets of Ω which satisfy that $\forall s \in [1, K]$, $\mathcal{D}_{\Omega_s} \geq \mathcal{D}_{\Omega}$. Then, following Theorem 4, we next describe a new extension of SVD. Note that, extensions to other LRMA methods can be similarly derived.

$$\begin{aligned} \hat{R} = \arg\min_{X} \ & \lambda_0 \mathcal{D}_{\Omega}(X) + \sum_{s=1}^{K} \lambda_s \mathcal{D}_{\Omega_s}(X) \\ s.t. \ & rank(X) = r. \end{aligned} \tag{3}$$

where $\lambda_0, \lambda_1, ..., \lambda_K$ define the contributions of each component in the loss function.

3.2. Hard Predictable Subsets Selection

The key step in Equation 3 is to obtain subsets of Ω — $\{\Omega_1, ..., \Omega_K\}$ which satisfy that $\forall s \in [1, K]$, $\mathcal{D}_{\Omega_s} \geq \mathcal{D}_{\Omega}$. To obtain such Ω_s is not trivial, because we can only check if the condition is satisfied with the final model. But the final model cannot be known before we define and optimize a given loss function. Here, we address this issue using the following idea: 1) approximate the targeted matrix R with existing LRMA solutions, e.g., RSVD (Paterek, 2007); 2) for each entry $(i,j) \in \Omega$, it is chosen with large probability if $|R_{i,j} - \hat{R}_{i,j}| < \mathcal{D}_{\Omega}$ and small probability otherwise; and 3) obtain Ω' by removing the chosen entries to satisfy the condition of Proposition 1, or probe Ω' to find hard predictable subsets that satisfy the condition of Theorem 4. By assuming that other LRMA methods will not dramatically differ from the final model of SMA, we can ensure that Ω' will satisfy $\mathcal{D}_{\Omega'} \geq \mathcal{D}_{\Omega}$ with high probability.

3.3. The SMA Learning Algorithm

The pseudo-code of the proposed SMA learning algorithm to solve the optimization problem defined in Equation 3 is presented in Algorithm 1. From Step 1 to 9, we obtain K different hard predictable entry sets. In Step 10, the optimization is performed by stochastic gradient descent, the details of which are trivial and thus omitted. Also, L_2 regularization is adopted in Step 10. Note that, other types of optimization methods and regularization can also be used in Algorithm 1. The complexity of Step 1 to 9 is $O(|\Omega|)$, where $|\Omega|$ is the number of the observed entries in R. The complexity of Step 10 is $O(rmn)$ per-iteration, where r is the rank and m, n is the matrix size. Thus, the computation complexity of SMA is similar to classic LRMA methods, such as regularized SVD (Paterek, 2007).

Algorithm 1 The SMA Learning Algorithm

```
Require: $R$ is the targeted matrix, $\Omega$ is the set of entries in $R$, and $\hat{R}$ is an approximation of $R$ by existing L-RMA methods. $p > 0.5$ is the predefined probability for entry selection. $\mu_1$ and $\mu_2$ are the coefficients for L2-regularization.
$\Omega' = \emptyset$;
for each $(i, j) \in \Omega$ do
    randomly generate $\rho \in [0, 1]$;
    if ($|R_{i,j} - \hat{R}_{i,j}| \leq \mathcal{D}_\Omega$ & $\rho \leq p$) or ($|R_{i,j} - \hat{R}_{i,j}| > \mathcal{D}_\Omega$ & $\rho \leq 1 - p$) then
        $\Omega' \leftarrow \Omega' \cup \{(i, j)\}$;
    end if
end for
randomly divide $\Omega'$ into $\omega_1, ..., \omega_K$ ($\cup_{k=1}^{K} \omega_i = \Omega'$);
for all $k \in [1, K]$, $\Omega_k = \Omega - \omega_k$;
$(\hat{U}, \hat{V}) := \arg\min_{U,V} [\sum_{k=1}^{K} \lambda_k \mathcal{D}_{\Omega_k}(U^T V)$
    $+\lambda_0 \mathcal{D}_\Omega(UV^T) + \mu_1 \parallel U \parallel^2 + \mu_2 \parallel V \parallel^2]$
return $\hat{R} = \hat{U}\hat{V}^T$
```

4. Experiments

In this section, we first analyze the generalization performance of SMA, and then evaluate the performance of S-MA with different parameters, e.g., rank r and the number of non-overlapping subsets K. Next, SMA is compared against seven state-of-then-art matrix approximation based recommendation algorithms, including four single MA methods and three ensemble methods. At last, we analyze SMA's accuracy in different data sparsity settings.

4.1. Experiment Setup

Two widely used datasets are adopted to evaluate SMA: MovieLens 10M (~70k users, 10k items, 10^7 ratings) and Netflix (~480k users, 18k items, 10^8 ratings). For each dataset, we randomly split it into training and test sets and keep the ratio of training set to test set as 9:1. All experimental results are presented by averaging the results over five different random train-test splits.

In this study, we use learning rate $v = 0.001$ for stochastic gradient decent method, $\mu_1 = 0.06$ for L_2-regularization coefficient, $\epsilon = 0.0001$ for gradient descent convergence threshold, and $T = 250$ for maximum number of iterations. Optimal parameters of the compared methods are chosen from their original papers. The source codes of all the experiments are publicly available [1].

We compare the performance of SMA with four single MA models and three ensemble MA models as follows:

- Regularized SVD [Paterek et al., KDD' 07]: is one of the most widely used matrix factorization methods, in which user/item features are estimated by minimizing the sum-squared error using L_2 regularization.
- BPMF [Salakhutdinov et al., ICML' 08]: is a Bayesian extension of PMF with model parameters and hyperparameters estimated using Markov chain Monte Carlo method.
- APG [Toh et al., PJO' 2010]: computes the approximation by solving a nuclear norm regularized linear least squares problem.
- GSMF [Yuan et al., AAAI' 14]: can transfer information among multiple types of user behaviors by modeling the shared and private latent factors with group sparsity regularization.
- DFC [Mackey et al., NIPS' 11]: is an ensemble method, which divides a large-scale matrix factorization task into smaller subproblems, solves each other in parallel, and finally combines the subproblem solutions.
- LLORMA [Lee et al., ICML' 13]: is an ensemble method, which assumes that the original matrix is described by multiple low-rank submatrices constructed by non-parametric kernel smoothing techniques.
- WEMAREC [Chen et al., SIGIR' 15]: is an ensemble method, which constructs biased model by weighting strategy to address the insufficient data issue in each submatrix.

4.2. Generalization Performance

Figure 2 compares training/test errors of SMA and RSVD with different epochs on MovieLens 10M dataset (rank $r = 20$ and subset number $K = 3$). As we can see, the differences between training and test error of SMA are much smaller than RSVD. Moreover, the training error and test error are very close when epoch is less than 100. This result demonstrates that SMA can indeed find models that have good generalization performance and yield small generalization error during the training process.

4.3. Sensitivity Analysis

Figure 3 investigates how SMA performs by varying number of non-overlapping subsets K (rank $r = 200$) and the optimal RMSEs of all compared methods on both Movielens 10M (left) and Netfilx (right) datasets. As we can see, SMA outperforms all these state-of-the-art methods with K varying from 1 to 5. It should be noted that, when $K = 0$, SMA is degraded to RSVD. Thus, the fact that S-MA can produce better recommendations than RSVD confirms Theorem 1: with additional terms $\sum_{s=1}^{K} \lambda_s \mathcal{D}_{\Omega_s}(\hat{R})$, we can improve the stability of MA models. In addition, we can see the RMSEs on both two datasets decrease as K increases. This further confirms Theorem 4: probing easily predictable entries to form harder predictable entry sets can better increase the model performance.

[1] https://github.com/ldscc/StableMA.git.

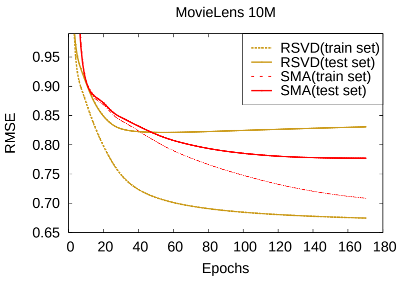

Figure 2. Training and test errors vs. epochs of RSVD and SMA on MovieLens 10M dataset.

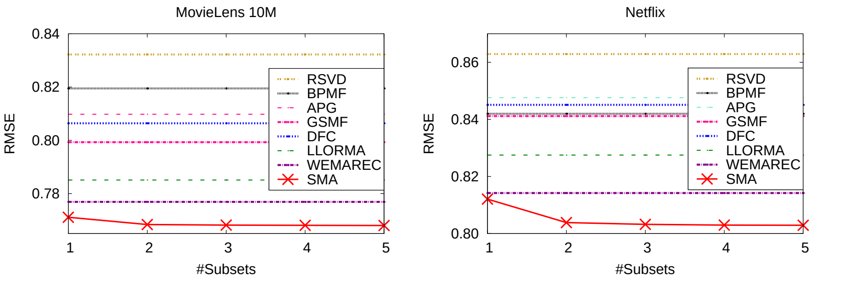

Figure 3. Effect of subset number K on MovieLens 10M dataset (left) and Netflix dataset (right). SMA models are indicated by solid line and other compared methods are indicated by dot lines.

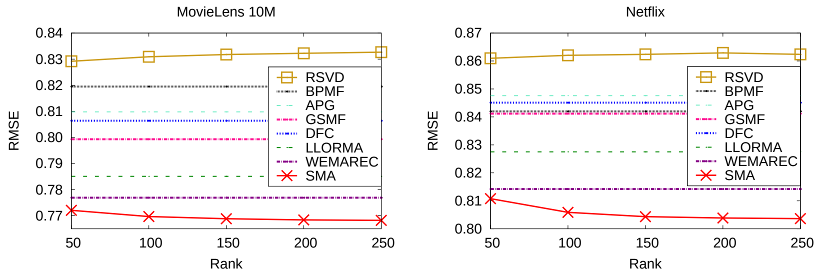

Figure 4. Effect of rank r on MovieLens 10M dataset (left) and Netflix dataset (right). SMA and RSVD models are indicated by solid line and other compared methods are indicated by dot lines.

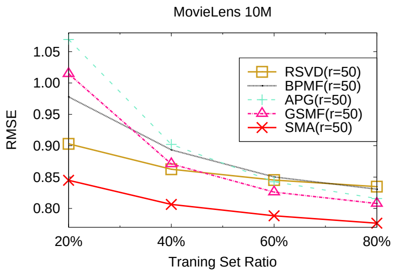

Figure 5. RMSEs of SMA and four single methods with varying training set size on MovieLens 10M dataset (rank $r = 50$).

Figure 4 analyzes the effect of rank r on MovieLens 10M (left) and Netflix (right) datasets by fixing $K = 3$. It can be seen that for any rank r from 50 to 250, SMA always outperform the other seven compared methods in recommendation accuracy. And higher ranks for SMA will lead to better accuracy when the rank r increases from 50 to 250 on both two datasets. It is interesting to see that the recommendation accuracies of RSVD decrease slightly when $r > 50$ due to over-fitting and SMA can consistently increase recommendation accuracy even when $r > 200$. This indicates that SMA is less prone to over-fitting than RSVD, i.e., SMA is more stable than RSVD.

Table 1. RMSEs of SMA and the seven compared methods on MovieLens (10M) and Netflix datasets.

	MovieLens (10M)	Netflix
RSVD	0.8256 ± 0.0006	0.8534 ± 0.0001
BPMF	0.8197 ± 0.0004	0.8421 ± 0.0002
APG	0.8101 ± 0.0003	0.8476 ± 0.0003
GSMF	0.8012 ± 0.0011	0.8420 ± 0.0006
DFC	0.8067 ± 0.0002	0.8453 ± 0.0003
LLORMA	0.7855 ± 0.0002	0.8275 ± 0.0004
WEMAREC	0.7775 ± 0.0007	0.8143 ± 0.0001
SMA	**0.7682 ± 0.0003**	**0.8036 ± 0.0004**

4.4. Accuracy Comparisons

Table 1 presents the performance of SMA with rank $r = 200$ and subset number $K = 3$. The compared methods are as follows: RSVD ($r = 50$) (Paterek, 2007), BPMF ($r = 300$) (Salakhutdinov & Mnih, 2008), APG ($r = 100$) (Toh & Yun, 2010), GSMF ($r = 20$) (Yuan et al., 2014), DFC ($r = 30$) (Mackey et al., 2011), LLORMA ($r = 20$) (Lee et al., 2013) and WEMAREC ($r = 20$) (Chen et al., 2015) on MovieLens 10M and Netflix datasets. Notably, DFC, LLORMA and WEMAREC are ensemble methods, which have been shown to be more accurate than single methods due to better generalization performance. However, as shown in Table 1, the SMA method significantly outperforms all seven compared methods on both two datasets. This confirms that SMA can indeed achieve better generalization performance than both state-of-the-art single methods and ensemble methods. The main reason is that SMA can minimize objective functions that lead to solutions with good generalization performance, but other methods cannot guarantee low gap between training error and test error.

4.5. Performance under Data Sparsity

Figure 5 presents the RMSEs of SMA vs. the size of training set size as compared with four single LRMA methods (RSVD, BPMF, APG and GSMF). The rank r of all five methods are fixed to 50. Note that, the rating density becomes more sparse when the training set ratio decreases. The results show that all methods can improve accuracy with the training set size increasing, but the proposed SMA method always outperforms the compared methods. This demonstrates that SMA can still provide stable matrix approximation even on very sparse dataset.

5. Related Work

Algorithmic stability has been analyzed and applied in several popular problems, such as regression (Bousquet & Elisseeff, 2001), classification (Bousquet & Elisseeff, 2001), ranking (Lan et al., 2008), marginal inference (London et al., 2013), etc. Bousquet & Elisseeff (2001) first proposed a method of obtaining bounds on generalization errors of learning algorithms, and formally proved that regularization networks posses the uniform stability property. Then, Bousquet & Elisseeff (2002) extends the algorithmic stability concept from regression to classification. Kutin & Niyogi (2002) generalized the work of Bousquet & Elisseeff (2001) and proposed the notion of training stability, which can ensure good generalization error bounds even when the learner has infinite VC dimension. Lan et al. (2008) proposed query-level stability and gave query-level generalization bounds to learning to rank algorithms. Agarwal & Niyogi (2009) derived generalization bounds for ranking algorithms that have good properties of algorithmic stability. Shalev-Shwartz et al. (2010) considered the general learning setting including most statistical learning problems as special cases, and identified that stability is the necessary and sufficient condition for learnability. London et al. (2013) proposed the concept of collective stability for structure prediction, and established generalization bounds for structured prediction. This work differs from the above works by (1) this work introduces the stability concept to low-rank matrix approximation problem, and proves that matrix approximations with high stability will have high probability to generalize well and (2) most existing works focus on theoretical analysis, but this work provides practical framework for achieving solutions with high stability.

Low-rank matrix approximation methods have been extensively studied recently. Lee & Seung (2001) analyzed the optimization problems of Non-negative Matrix Factorization (NMF). Srebro et al. (2004b) proposed Maximum-Margin Matrix Factorization (MMMF), which can learn low-norm factorizations by solving a semi-definite program to achieve collaborative prediction. Salakhutdinov & Mnih (2007) viewed matrix factorization from a probabilistic perspective and proposed Probabilistic Matrix Factorization (PMF). Later, they proposed Bayesian Probabilistic Matrix Factorization (BPMF) (Salakhutdinov & Mnih, 2008) by giving a fully Bayesian treatment to PMF. Lawrence & Urtasun (2009) also extends PMF and developed a non-linear PMF using Gaussian process latent variable models. Paterek (2007) applied regularized singular value decomposition (RSVD) in the Netflix Prize contest. Koren (2008) combined matrix factorization and neighborhood model and built a more accurate combined model named SVD++. Many of the above methods tried to solve overfitting problems in model training, e.g., regularization in most of the above methods and Bayesian treatment in BPMF. However, alleviating overfitting cannot decrease the lower bound of generalization errors, and thus cannot fundamentally solve the low generalization performance problem. Different from the above works, this work proposes a new optimization problem with smaller lower bound of generalization error. Minimizing the new loss function can substantially improve generalization performance of matrix approximation as demonstrated in the experiments. Srebro et al. (2004a) analyzed the generalization error bounds of collaborative prediction with low-rank matrix approximation for “0-1” recommendation. Candès & Plan (2010) established error bounds of matrix completion problem with noises. However, those works did not consider how to achieve LRMA with small generalization error.

Ensemble matrix approximation methods, such as ensemble MMMF (DeCoste, 2006), DFC (Mackey et al., 2011), LLORMA (Lee et al., 2013), WEMAREC (Chen et al., 2015), ACCAMS (Beutel et al., 2015) etc., have been proposed, which aimed to provide matrix approximations with high generalization performance by ensemble learning. However, those ensemble methods need to train a number of biased weak matrix approximation models, which require much more computations than SMA. In addition, weak models in those methods are generated by heuristics which are not directly related to minimizing generalization error. Therefore, the optimality of generalization performance of those methods cannot be proved as in this work.

6. Conclusion

Low-rank matrix approximation methods are widely adopted in machine learning applications. However, similar to other machine learning techniques, many existing low-rank matrix approximation methods suffer from the low generalization performance issue. This paper introduces the stability notion to low-rank matrix approximation problem, in which models achieve high stability will have better generalization performance. Then, SMA, a new low-rank matrix approximation framework, is proposed to achieve high stability, i.e., high generalization performance. Experimental results on real-world datasets demonstrate that the proposed SMA method can achieve better prediction accuracy than both state-of-the-art matrix approximation methods and ensemble methods in recommendation task.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61233016, and the National Science Foundation of USA under Grant Nos. 0954157, 1251257, 1334351, and 1442971.

References

Agarwal, Shivani and Niyogi, Partha. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10:441–474, 2009.

Beutel, Alex, Ahmed, Amr, and Smola, Alexander J. ACCAM-S: additive co-clustering to approximate matrices succinctly. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 119–129, 2015.

Bousquet, Olivier and Elisseeff, André. Algorithmic stability and generalization performance. In *Advances in Neural Information Processing Systems*, pp. 196–202, 2001.

Bousquet, Olivier and Elisseeff, André. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Candès, Emmanuel J. and Plan, Yaniv. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Candès, Emmanuel J. and Recht, Benjamin. Exact matrix completion via convex optimization. *Communications of ACM*, 55 (6):111–119, 2012.

Chen, Chao, Li, Dongsheng, Zhao, Yingying, Lv, Qin, and Shang, Li. WEMAREC: Accurate and scalable recommendation through weighted and ensemble matrix approximation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 303–312, 2015.

DeCoste, Dennis. Collaborative prediction using ensembles of maximum margin matrix factorizations. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 249–256, 2006.

Keshavan, Raghunandan H., Montanari, Andrea, and Oh, Sewoong. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

Kohavi, Ron. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1137–1145, 1995.

Koren, Yehuda. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 426–434, 2008.

Koren, Yehuda, Bell, Robert, and Volinsky, Chris. Matrix factorization techniques for recommender systems. *Computer*, 42 (8):30–37, 2009.

Kutin, Samuel and Niyogi, Partha. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence*, pp. 275–282, 2002.

Lan, Yanyan, Liu, Tie-Yan, Qin, Tao, Ma, Zhiming, and Li, Hang. Query-level stability and generalization in learning to rank. In *Proceedings of the 25th international conference on Machine learning*, pp. 512–519, 2008.

Lawrence, Neil D. and Urtasun, Raquel. Non-linear matrix factorization with gaussian processes. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 601–608, 2009.

Lee, Daniel D and Seung, H Sebastian. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 556–562, 2001.

Lee, Joonseok, Kim, Seungyeon, Lebanon, Guy, and Singer, Y-oram. Local low-rank matrix approximation. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 82–90, 2013.

London, Ben, Huang, Bert, Taskar, Ben, and Getoor, Lise. Collective stability in structured prediction: Generalization from one example. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 828–836, 2013.

Mackey, Lester W, Jordan, Michael I, and Talwalkar, Ameet. Divide-and-conquer matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 1134–1142, 2011.

Paterek, Arkadiusz. Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD cup and workshop*, volume 2007, pp. 5–8, 2007.

Salakhutdinov, Ruslan and Mnih, Andriy. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 1257–1264, 2007.

Salakhutdinov, Ruslan and Mnih, Andriy. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, pp. 880–887. ACM, 2008.

Shalev-Shwartz, Shai, Shamir, Ohad, Srebro, Nathan, and Sridharan, Karthik. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

Srebro, Nathan, Alon, Noga, and Jaakkola, Tommi S. Generalization error bounds for collaborative prediction with low-rank matrices. In *Advances in Neural Information Processing Systems*, pp. 1321–1328, 2004a.

Srebro, Nathan, Rennie, Jason D. M., and Jaakkola, Tommi S. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 1329–1336, 2004b.

Toh, Kim-Chuan and Yun, Sangwoon. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640): 15, 2010.

Yan, Junchi, Zhu, Mengyuan, Liu, Huanxi, and Liu, Yuncai. Visual saliency detection via sparsity pursuit. *IEEE Signal Processing Letters*, 17(8):739–742, 2010.

Yuan, Ting, Cheng, Jian, Zhang, Xi, Qiu, Shuang, and Lu, Hanqing. Recommendation by mining multiple user behaviors with group sparsity. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pp. 222–228, 2014.