Low-Rank Matrix Approximation with Stability

Dongsheng Li[1], Chao Chen[2], Qin (Christine) Lv[3], Junchi Yan[1], Li Shang[3], Stephen M. Chu[1]

[1]IBM Research - China, [2]Tongji University, [3]University of Colorado Boulder

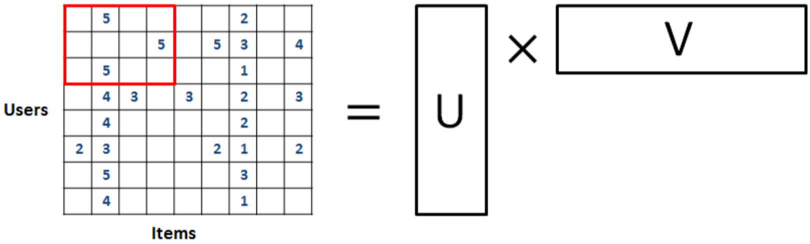
Low-Rank Matrix Approximation (LRMA)

$$U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}, \ s.t. \ \hat{R} = UV^T$$

The optimization problem of LRMA can be described as follows:

$$\hat{R} = \arg\min_X Loss(R, X), \ s.t. \ rank(X) = r$$

Example: User-item ratings matrix used by recommender systems



Problem Formulation

Generalization performance is a problem of matrix approximation when data is sparse, incomplete, and noisy [Keshavan et al., 2010; Candès & Recht, 2012].

- models are biased to the limited training data (sparse, incomplete)
- small changes in the training data (noisy) may significantly change the models.

Algorithmic stability has been introduced to investigate the generalization error bounds of learning algorithms [Bousquet & Elisseeff, 2001; 2002]. A stable learning algorithm has the properties that

- slightly changing the training set does not result in significant change to the output
- the training error should have small variance
- the training errors are close to the test errors

Stability w.r.t Matrix Approximation

Definition (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:

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

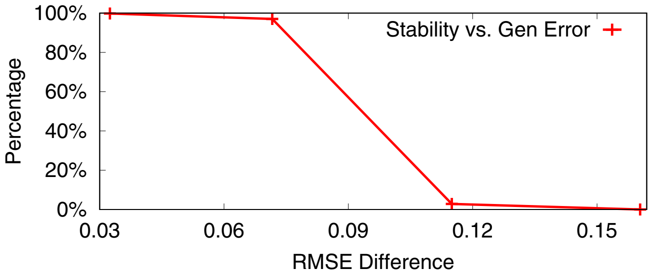


Figure: Stability vs. generalization error of RSVD on the MovieLens (1M) dataset. Rank $r = 5, 10, 15, 20$ and $\epsilon = 0.0046$. 500 runs.

Theoretical Analysis

Theorem

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 $\delta_1 \leq \delta_2$.

Remark

1. *If we select a subset of entries Ω' from Ωthat are harder to predict than average, then minimizing $\lambda_0 \mathcal{D}_\Omega(\hat{R}) + \lambda_1 \mathcal{D}_{\Omega'}(\hat{R})$ will be more stable than minimizing $\mathcal{D}_\Omega(\hat{R})$.*

Theoretical Analysis

Theorem

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* $\delta_1 \leq \delta_2$*.*

Remark

2. *Removing more entries that are easy to predict will yield more stable matrix approximation.*

Theorem

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* $\delta_1 \leq \delta_2$*.*

Remark

3. *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.*

New Optimization Problem

We propose the SMA (Stable MA) framework that is generally applicable to any LRMA methods.

E.g., a new extension of SVD:

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

where $\lambda_0, \lambda_1, ..., \lambda_K$ define the contributions of each component in the loss function. (Extensions to other LRMA methods can be similarly derived.)

The SMA Learning Algorithm

Require: R is the targeted matrix, Ω is the set of entries in R, and $\hat{R}$ is an approximation of R by existing LRMA methods. $p > 0.5$ is the predefined probability for entry selection. μ_1 and μ_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 Ω' into $\omega_1, ..., \omega_K$ $(\cup_{k=1}^{K} \omega_i = \Omega')$;
for all $k \in [1, K]$, $\Omega_k = \Omega - \omega_k$;
$(\hat{U}, \hat{V}) := \arg\min_{U,V} [\sum_{k=1}^{K} \lambda_k \mathcal{D}_{\Omega_k}(UV^T)$
$+\lambda_0 \mathcal{D}_\Omega(UV^T) + \mu_1 \parallel U \parallel^2 + \mu_2 \parallel V \parallel^2]$
return $\hat{R} = \hat{U}\hat{V}^T$

Datasets

- **MovieLens 10M** ($\sim$70k users, 10k items, 10^7 ratings)
- **Netflix** ($\sim$480k users, 18k items, 10^8 ratings)

Performance comparison with four single MA models and three ensemble MA models as follows:

- Regularized SVD [Paterek et al., KDD' 07].
- BPMF [Salakhutdinov et al., ICML' 08].
- APG [Toh et al., PJO' 2010].
- GSMF [Yuan et al., AAAI' 14].
- DFC [Mackey et al., NIPS' 11].
- LLORMA [Lee et al., ICML' 13].
- WEMAREC [Our prior work, SIGIR' 15].

Generalization Performance

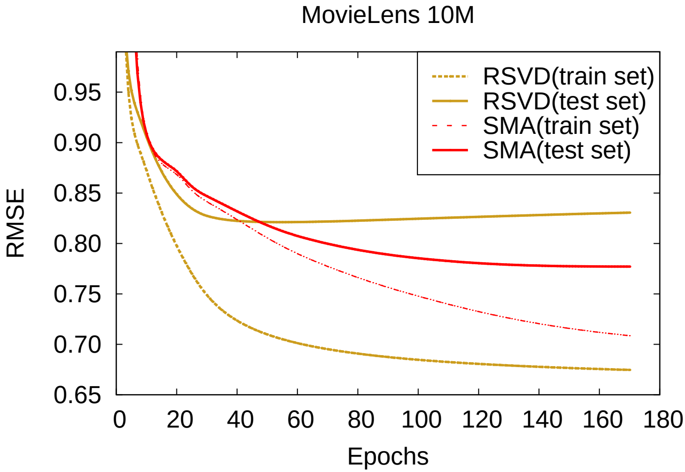


Figure: Training and test errors vs. epochs of RSVD and SMA on the MovieLens 10M dataset.

Sensitivity of Subset Number K

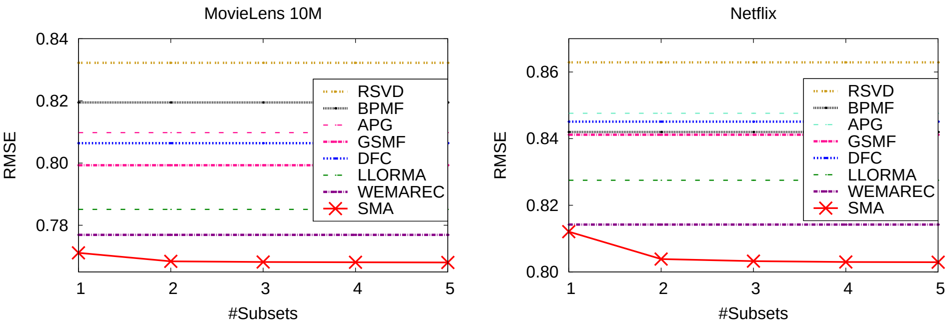


Figure: Effect of subset number K on MovieLens 10M dataset (left) and Netflix dataset (right). SMA and RSVD models are indicated by solid lines and other compared methods are indicated by dotted lines.

Sensitivity of Rank r

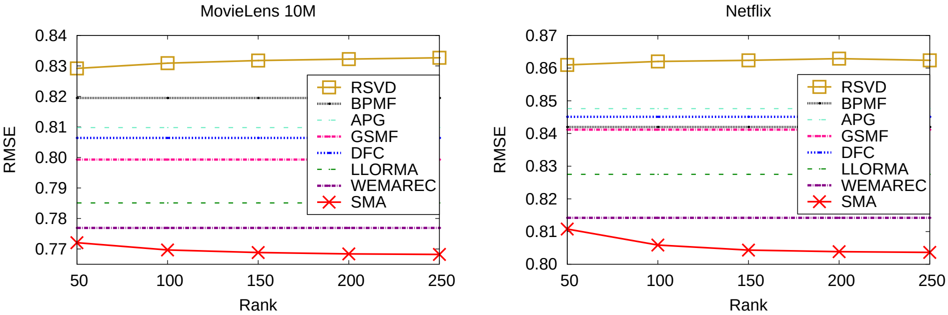


Figure: Effect of rank r on MovieLens 10M dataset (left) and Netflix dataset (right). SMA and RSVD models are indicated by solid lines and other compared methods are indicated by dotted lines.

Sensitivity of Training Set Size

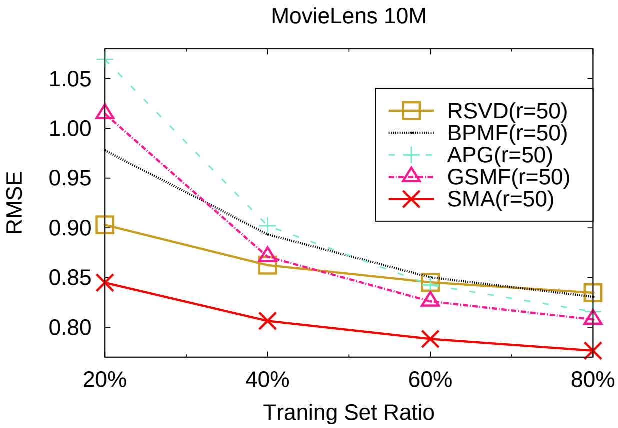


Figure: RMSEs of SMA and four single methods with varying training set size on MovieLens 10M dataset (rank $r = 50$).

Table: RMSE Comparison of SMA and Seven Other Methods

	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	$\mathbf{0.7682 \pm 0.0003}$	$\mathbf{0.8036 \pm 0.0004}$

Conclusion

SMA (Stable MA), a new low-rank matrix approximation framework, is proposed, which can

- achieve high stability, i.e., high generalization performance;
- achieve better accuracy than state-of-the-art MA-based CF methods;
- achieve good accuracy with very sparse datasets.

Source code available at:
https://github.com/ldscc/StableMA.git