

RESEARCH

Open Access



Nonconvex matrix completion with Nesterov's acceleration

Xiao-Bo Jin^{1*}, Guo-Sen Xie², Qiu-Feng Wang³, Guoqiang Zhong⁴ and Guang-Gang Geng⁵

*Correspondence:

xbjin9801@gmail.com

¹Henan University of Technology, Lianhua, Zhengzhou, China

Full list of author information is available at the end of the article

Abstract

Background: In matrix completion fields, the traditional convex regularization may fall short of delivering reliable low-rank estimators with good prediction performance. Previous works use the alternation least squares algorithm to optimize the nonconvex regularization. However, this algorithm has high time complexities and requires more iterations to reach convergence, which cannot scale to large-scale matrix completion problems. We need to develop faster algorithm to examine large amount of data to uncover the correlations between items for the big data analytics.

Results: In this work, we adopt the randomized SVD decomposition and Nesterov's momentum to accelerate the optimization of nonconvex matrix completion. The experimental results on four real world recommendation system datasets show that the proposed algorithm achieves tremendous speed increases whilst maintaining comparable performance with the alternating least squares (ALS) method for nonconvex (convex) matrix completions.

Conclusions: Our novel algorithm can achieve comparable performance with previous algorithms but with shorter running time.

Keywords: Matrix completion, Nonconvex optimization, Randomized SVD, Nesterov's momentum

Background

The fields of machine learning and mathematical programming are increasingly intertwined. Optimization algorithms lie at the heart of most machine learning approaches [1] and have been widely applied in learning to rank [2, 3], clustering analysis [4], matrix factorization [5], etc.

Matrix completion is the task of filling in the missing entries of a partially observed matrix. Given a partially observed matrix O , matrix completion [6] attempts to recover a low-rank matrix X that approximates O on the observed entries. This approach leads to the following low-rank regularized optimization problem:

$$\min_X \frac{1}{2} \|P_\Omega(O - X)\|_F^2 + \lambda \cdot \text{rank}(X), \quad (1)$$

where P_Ω is the projection of $X_{m \times n}$ onto the observed indices Ω and otherwise zeros, and $\|\cdot\|_F$ represents the Frobenius norm of a matrix.



Direct minimization of the matrix rank is NP-hard, and a nature convexification of $\text{rank}(X)$ is the nuclear norm $\|X\|_*$ of X . The corresponding surrogate problem can be solved by various optimization algorithms. Candes and Recht [6] realized that convex relaxation is the equivalent to classic semidefinite programming (SDP), where SDP complexity is $O(\max(m, n)^4)$. Cai et al. [7] gave a first-order method that approximately solves convex relaxation by the singular value thresholding algorithm (SVT), which takes $O(mn^2)$ time to solve SVD in each SVT iteration. The soft-impute algorithm utilized a special “sparse + low-rank” structure associated with SVT to efficiently compute the SVD with a reasonable time. The augmented Lagrangian method (ALM) provides a powerful framework to solve convex problems with equality constraints. Lin et al. [8] used ALM to solve matrix completion problems. The fixed point continuation with approximate SVD (FPCA) algorithm [9] is based on the proximal gradient with a continuation technique to accelerate convergence. Toh and Yun [10] proposed an accelerated proximal gradient algorithm on the balanced approach. Tomioka et al. [11] used a dual augmented Lagrangian algorithm for matrix completion, which achieves super-linear convergence and can be generalized to the problem of learning multiple matrices and general spectral regularization.

The traditional nuclear norm regularization may fall short of delivering reliable low-rank estimators with good prediction performance. Many researchers move their strides beyond the convex l_1 -penalty to the nonconvex form. Mazumder et al. [12] used MC+penalty to demonstrate the performance of the coordinate-descent algorithm. Dimitris et al. [13] developed a discrete extension of modern first-order continuous optimization methods to find a high-quality feasible solution. Zhang [14] proposed the MC+ method for penalized variable selection in high-dimensional linear regression, which is brought into a unified view of concave high-dimensional sparse estimation procedures [15].

Previous work [16] used the alternating least squares (ALS) procedure to optimize the nonconvex matrix completion. However, this method has high time complexities and requires more iterations to reach convergence, which cannot meet the requirements of big data analytics. Halko et al. [17] presented a modular framework for constructing randomized algorithms to compute the partial matrix decompositions. Nesterov’s momentum [18] makes a step in the velocity direction and then makes a correction to a velocity vector based on the new location. Nesterov’s momentum has been widely used in deep learning [19] and signal processing [20].

In this work, we adopt the randomized SVD decomposition and Nesterov’s momentum to accelerate optimization of nonconvex matrix completion. Randomized SVD decomposition requires very few iterations to converge quickly. In addition, the proposed algorithm has a low running time and fast convergence rate by Nesterov’s acceleration. The experimental results on four real world recommendation system datasets show that the proposed algorithm achieves tremendous speed increases whilst maintaining comparable performance with ALS for nonconvex (convex) matrix completions.

Our contribution is as follows: (1) we introduce Nesterov’s momentum to solve the nonconvex matrix completion problem; (2) we further adopt the random SVD decomposition to accelerate the converge of the algorithm; (3) the experimental results on benchmark datasets shows the effectiveness of our algorithm.

The remainder of the paper is organized as follows: the “Background” section introduces nonconvex matrix completion and the alternating least squares algorithm; the “Nonconvex Matrix Completion by Nesterov’s Acceleration (NMCNA)” section shows the experimental results; the “Discussion” section analyses the time complexities of the algorithms; the “Conclusionss” section summarizes the conclusions; and the “Methods” gives our novel algorithm called Nonconvex Matrix Completion by Nesterov’s Acceleration (NMCNA).

Nonconvex Matrix Completion

Nonconvex matrix completion uses nonconvex regularizers to approximate the l_0 -penalty, which leads to the following nonconvex estimation problems:

$$\min_X \frac{1}{2} \|P_\Omega(X - Y)\|_F^2 + \sum_{i=1}^{\min\{m,n\}} P(\sigma_i(X); \lambda, \gamma), \tag{2}$$

where $\sigma_i(X) (i \geq 1)$ are the singular values of X and $P(\sigma_i(X); \lambda, \gamma)$ is a concave penalty function of σ on $[0, \infty]$. The parameter (λ, γ) controls the amount of nonconvexity and shrinkage. For the nuclear norm regularized problem with $P(\sigma_i(X); \lambda, \gamma) = \lambda\sigma_i(X)$, the soft-thresholding operator gives a solution of problem (2):

$$S_\lambda(Z) = U \text{diag}(s_\lambda(\sigma)) V^T, \quad (s_\lambda(\sigma))_i = (\sigma_i - \lambda)_+, \tag{3}$$

where U and V is the left singular matrix and the right singular matrix of Z , σ is a singular value vector of Z and λ is a threshold.

For the more general spectral penalty function $P(\sigma; \lambda, \gamma)$, we obtain the following similar result [16] as Lemma 1.

Lemma 1 *Let $Z = U \text{diag}(\sigma) V^T$ denote the SVD of Z and $D_{\lambda,\gamma}$ denote the following thresholding operator on the singular values of Z :*

$$s_{\lambda,\gamma}(\sigma) \in \operatorname{argmin}_{\alpha \geq 0} \sum_{i=1}^{\min\{m,n\}} \frac{1}{2} (\alpha_i - \sigma_i)^2 + P(\alpha_i; \lambda, \gamma). \tag{4}$$

Then, $S_{\lambda,\gamma}(Z) = U \text{diag}(s_{\lambda,\gamma}(\sigma)) V^T$.

By the separability of the optimization, we can see that it can be converted into the optimization of the single variable

$$s_{\lambda,\gamma}(\sigma_i) \in \operatorname{argmin}_{\alpha \geq 0} \left\{ \frac{1}{2} (\alpha - \sigma_i)^2 + P(\alpha; \lambda, \gamma) \right\}. \tag{5}$$

The popular nonconvex function used in the high-dimensional regression framework includes the l_γ penalty, the SCAD penalty, the MC+ penalty and the log-penalty. In this work, we use the MC+ penalty ($\lambda > 0$ and $\gamma > 1$) as a representative:

$$P(\sigma; \lambda, \gamma) = \begin{cases} \lambda \left(\sigma - \frac{\sigma^2}{2\lambda\gamma} \right), & 0 \leq \sigma \leq \lambda\gamma, \\ \frac{\lambda^2\gamma}{2}, & \sigma \geq \lambda\gamma. \end{cases} \tag{6}$$

where the parameter λ and γ) controls the amount of nonconvexity and shrinkage.

We can easily analyze the solution of the thresholding function induced by the MC+ penalty:

$$s_{\lambda,\gamma}(\sigma) = \begin{cases} 0, & \sigma \leq \lambda, \\ \frac{\sigma-\lambda}{1-1/\gamma}, & \lambda < \sigma < \lambda\gamma, \\ \sigma, & \sigma > \lambda\gamma. \end{cases} \tag{7}$$

As observed in Fig. 1, when $\gamma \rightarrow +\infty$, the above threshold operator is consistent with the soft-thresholding operator $\lambda\sigma$. On the other hand, the threshold operator approaches the discontinuous hard-thresholding operator $\sigma 1(\sigma \geq \lambda)$ when $\gamma \rightarrow 1+$.

It is critical to solve the SVD efficiently. Mazumder [16] et al. used the alternating least square (ALS) procedure to compute a low-rank SVD.

Matrix Completion with Fast Alternating Least Squares

The ALS method [21, 22] solves the following nonlinear optimization problem:

$$\min_{A_{m \times r}, B_{n \times r}} \|P_{\Omega}(Y - AB^T)\|_F^2 + \lambda (\|A\|_F^2 + \|B\|_F^2), \tag{8}$$

where λ is a regularization parameter and $P_{\Omega}(X)$ is the projection of the matrix X on the matrix Ω . We call it as the maximum-margin matrix factorization (MMMF) criterion with nonconvex but biconvex. It can be solved with the alternating minimization algorithm. One variable is fixed (A or B), and we wish to solve (8) for another variable (B or A). We can easily see that this problem decouples into n separate ridge regressions. Further, we can prove that [21] the problem (8) is equivalent to the following problem:

$$\min_{X: \text{rank}(X) \leq r} \frac{1}{2} \|P_{\Omega}(Y - X)\|_F^2 + \lambda \|X\|_*, \tag{9}$$

where $X \approx AB$.

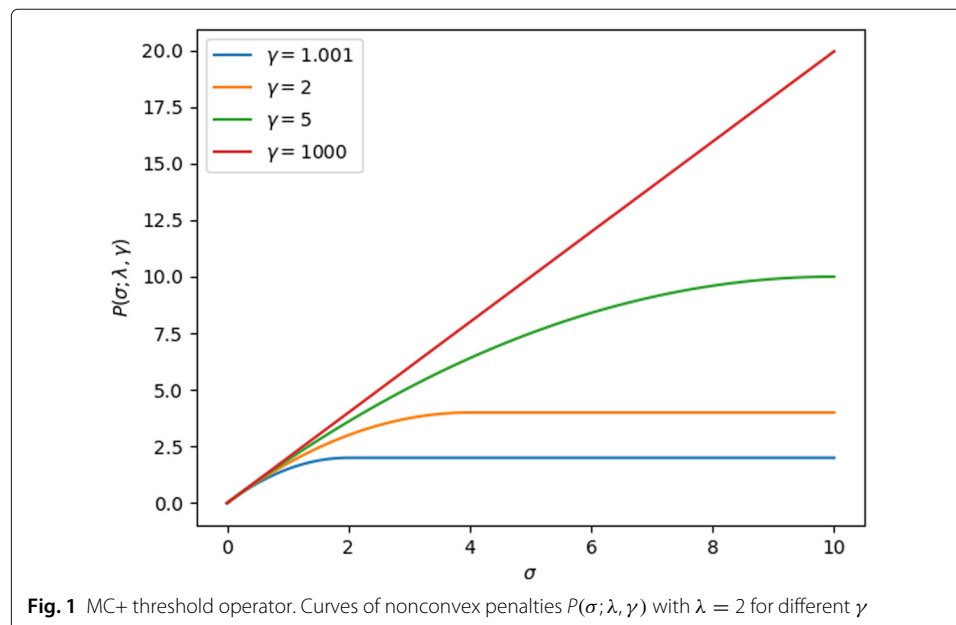


Fig. 1 MC+ threshold operator. Curves of nonconvex penalties $P(\sigma; \lambda, \gamma)$ with $\lambda = 2$ for different γ

Algorithm 1 Soft Impute with ALS

Choose randomly k columns of a $m \times m$ orthonormal matrix to construct $U_{m \times k}$

$V = 0, D = I_k, A = UD$

for (λ_i, γ_i) with the descending sequences $\{\lambda_i\}_{i=1}^t$ and $\{\gamma_i\}_{i=1}^t$ **do**

while ratio $> \epsilon$ **do**

$\tilde{U} = U, \tilde{V} = V, \tilde{D} = D$

 Solve (8) fixed A to obtain a solution:

$$\bar{B}^T = (D^2 + \lambda_i I)^{-1} D U^T \bar{X} \text{ with } \bar{X} = P_{\Omega} (Y - A B^T) + A B^T. \tag{10}$$

 Compute the orthonormal subspace spanned by \bar{B} :

$$\bar{B} D = U \tilde{D}^2 V^T. \tag{11}$$

$B = V \tilde{D}$

 Solve (8) fixed B to obtain a solution:

$$\bar{A}^T = (D^2 + \lambda_i I)^{-1} D V^T \bar{X}^T \text{ with } \bar{X}^T = P_{\Omega} (Y - B A^T) + B A^T. \tag{12}$$

 Compute the orthonormal subspace spanned by \bar{A} :

$$\bar{A} D = U \tilde{D}^2 V^T. \tag{13}$$

$A = U \tilde{D}$

$$\text{ration} = \frac{\|UD^2V^T - \tilde{U}\tilde{D}^2\tilde{V}^T\|_F^2}{\|UD^2V^T\|_F^2}$$

end while

$\bar{X} V = U D R^T$

$V = VR$

 Compute $D = \text{diag}(s_{\lambda_i, \gamma_i}(\sigma))$ with Eq. (7).

end for

It is noticed that (10) or (12) are efficient *sparse + low-rank* representations for high-dimensional problems, which can efficiently store and multiply the matrices. We can then obtain the analytical solution for the linear regression problems. Comparing (8) and (9), we can observe that $X = UD^2V^T = AB^T$ with $A = UD$ and $B = VD$.

Finally, we use the relative change in Frobenius norm to check for convergence

$$\begin{aligned} \text{ration} &= \frac{\|UD^2V^T - \tilde{U}\tilde{D}^2\tilde{V}^T\|_F^2}{\|UD^2V^T\|_F^2} \\ &= \frac{\text{tr}(D^4) + \text{tr}(\tilde{D}^4) - 2\text{tr}(D^2U^T\tilde{U}\tilde{D}^2\tilde{V}^TV)}{\text{tr}(D^4)} \end{aligned} \tag{14}$$

Results

In this section, we describe the experimental protocols and then compare the effectiveness and efficiency of our proposed algorithm (NMCNA) and the ALS-convex and

ALS-nonconvex algorithms. All algorithm are implemented in Python language with numpy library and running on Ubuntu system 14.04.

We used four real world recommendation system datasets (as shown in Table 1) from **ml100k** and **ml1m** provided by MovieLens¹ to compare nonconvex MC+ regularization. Every matrix item is taken from the ratings {1, 2, 3, 4, 5}.

For performance evaluation, we use root mean squared error as follows:

$$RMSE = \sqrt{\|P_{\Omega}(Y - O)\|_F^2 / \|\Omega\|_1}, \tag{15}$$

where Y and O are the true values and predicted ones, respectively.

For all three settings above, we fix $\lambda_1 = \|P_{\Omega}(Y)\|$ and $\lambda_{50} = 0.001 \cdot \lambda_1$. We choose 10 γ -values in a logarithmic grid from 10^6 to 1.1. We let the true ranks be taken from {3, 5, 10, 20, 30, 40, 50}.

From Fig. 2, we observed that a moderate γ value will obtain the best generalization performance.

Figure 3 shows the variations of the test RMSE with the training RMSE results. The x-axis is the training RMSE, and the y-axis the test RMSE in each subfigure. The rows and columns correspond to the ranks and the datasets.

Figure 4 shows that the running time varies with the iteration times. It is noted that our proposed algorithm runs faster than the ALS-convex and ALS-nonconvex algorithms.

Discussion

In Fig. 2, it is more obvious for the large rank $k = 50$ that the choice $\gamma = 80$ achieves the best test RMSE with 35 iterations for the ml-100k/ua dataset. Similar results also hold for the rank $k = 3$.

We can observe from Fig. 3 that the proposed algorithm can converge rapidly, although it has a high training and test RMSE from the beginning. In the end, comparable performance is achieved with ALS-convex and ALS-convex algorithms. The minimum RMSE value does not decrease with increasing ranks.

Figure 4 shows that our algorithm is **18 times** faster than other two algorithms on the ml-1m/ra.train and ml-1m/rb.train datasets.

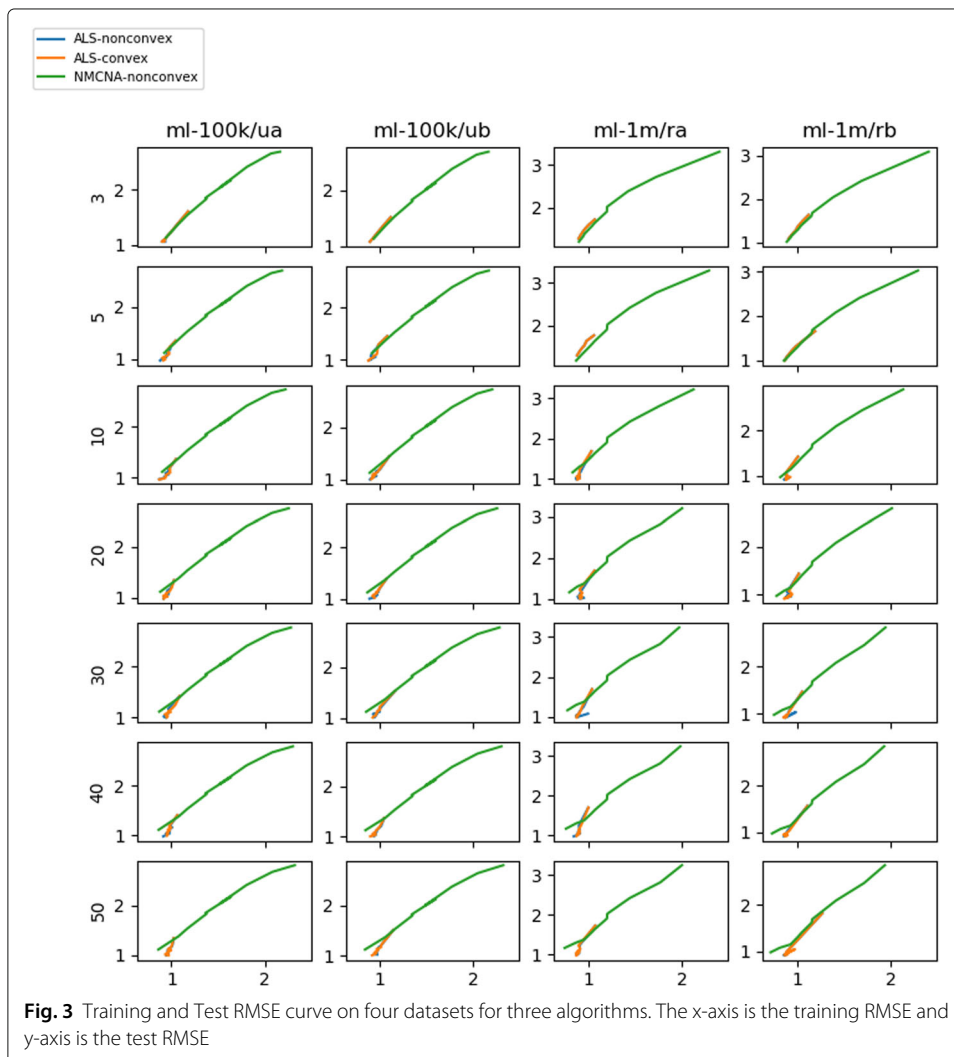
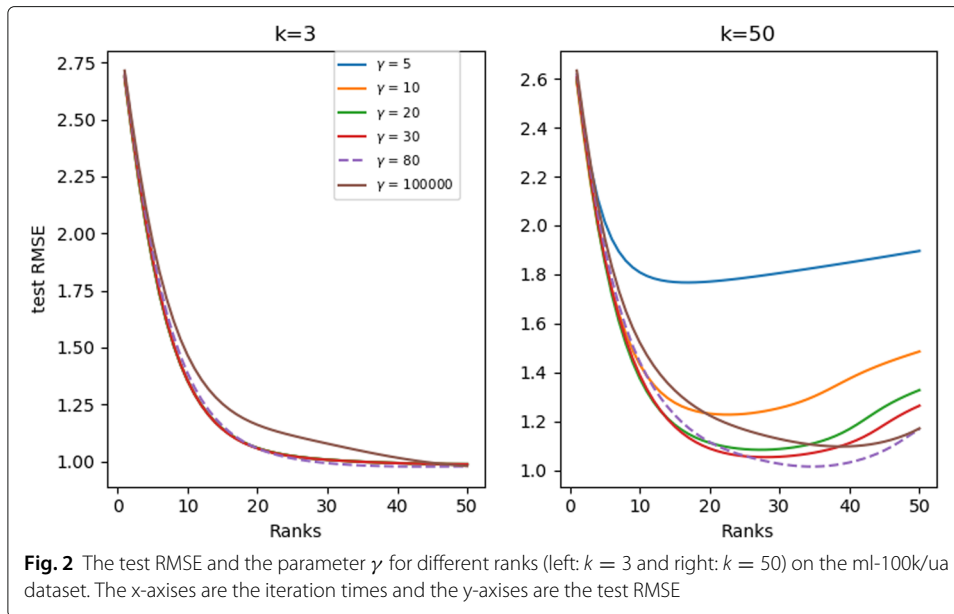
According to the discussion on the ALS algorithm in [21], the total cost of an iteration for computing SVD is $O(2r|\Omega| + mk^2 + 3nk^2 + k^3)$.

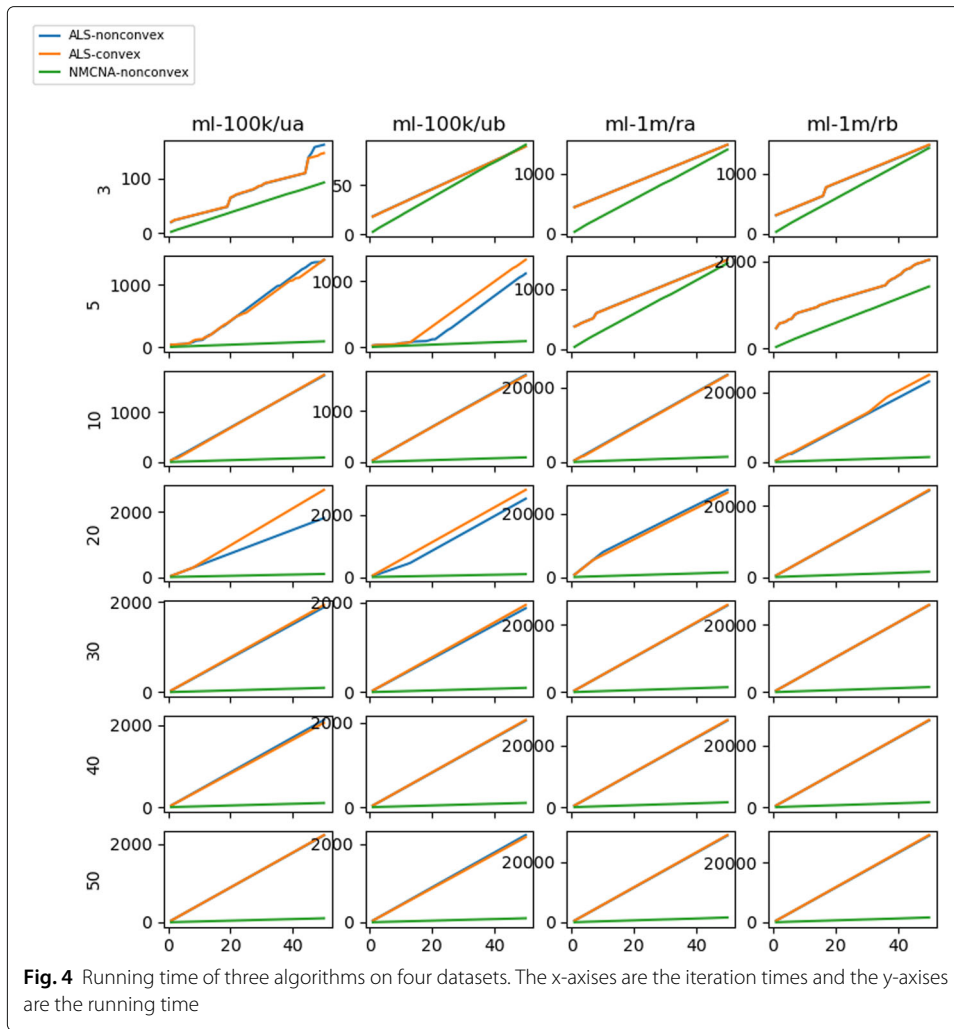
The randomized SVD procedure [17] in the proposed algorithm requires the computation of $Y^*Y^{*T}R$ (line 5 in Alg. 3), where $Y^{*T}R$ can be computed as

Table 1 Descriptions of the datasets

Dataset	#users	#movies	#(train items)	#(test items)
ml-100k/ua	943	1682	90,570	9430
ml-100k/ub	943	1682	90,570	9430
ml-1m/ra	6040	3952	939,809	60,400
ml-1m/rb	6040	3952	939,809	60,400

The details of the datasets used in experiments





$$\begin{aligned}
 Y^{*T}R &= P_{\Omega}(Y - X^*)R + (1 + \theta)XR - \theta X'R \\
 &= P_{\Omega}(Y - X^*)R + (1 + \theta)U\text{diag}(s_{\lambda,\gamma}(\sigma))V^T R \\
 &\quad - \theta U' \text{diag}(s'_{\lambda,\gamma}(\sigma))V'^T R
 \end{aligned} \tag{16}$$

Assume that X and X' have ranks r and r' ; then, it will take $O((r + r')|\Omega|)$ time for constructing $P_{\Omega}(Y - X^*)$ and $O(|\Omega|k)$ for computing $P_{\Omega}(Y - X^*)R$ time. Further, the cost of computing $U(\text{diag}(s_{\lambda,\gamma}(\sigma))(V^T R))$ is $O(mrk + nrk)$; then, the total cost is

$$O((r + r' + k)|\Omega| + (r' + r)(m + n)k). \tag{17}$$

Although we do not judge the performance advantages of random SVD over ALS for one iteration, the random SVD is shown to require fewer iterations (< 5) than ALS in these experiments.

Conclusions

This paper proposes a novel algorithm for nonconvex matrix completion. The proposed algorithm uses randomized SVD decomposition with few iterations. Nesterov’s acceleration further decreases its running time. The experimental results on four real-world recommendation system datasets show that the approach can accelerate the optimization of nonconvex matrix problems but with comparable performance to ALS algorithms. In the future, we will focus on the validation of other nonconvex functions for matrix completion problems.

Methods

Nonconvex Matrix Completion by Nesterov’s Acceleration (NMCNA)

Although soft-impute is a proximal gradient method that can be accelerated by Nesterov’s acceleration, the special *sparse+low-rank* structure will be lost in the ALS algorithm [23]. In this work, we resorted to the randomized partial matrix decomposition algorithm [17] to compute SVD.

Now, we present Nesterov’s acceleration for the proposed nonconvex matrix completion problem. First, define the following sequences:

$$c_i = \begin{cases} \frac{1+\sqrt{1+4c_{i-1}^2}}{2}, & i > 1, \\ 1, & i = 0, 1. \end{cases} \tag{18}$$

$$\theta_i = \frac{c_i - 1}{c_{i+1}} \tag{19}$$

Given the singular value decomposition of X and \bar{X} : $\bar{X} = U\Sigma V^T$ (Σ is diagonal) and $\bar{X}' = U'\Sigma'V'^T$, the proximal gradient algorithm uses the proximal mapping

$$X = U\text{diag}(s_{\lambda,\gamma}(\sigma))V^T, \tag{20}$$

$$X' = U'\text{diag}(s'_{\lambda,\gamma}(\sigma))V'^T, \tag{21}$$

where $s_{\lambda,\gamma}(\sigma)$ is computed by (7) and X (X') is the current (previous) predict matrix. By Nesterov’s acceleration, we have

$$X^* = (1 + \theta)X - \theta X' \tag{22}$$

$$Y^* = P_{\Omega}(Y - X^*) + X^*. \tag{23}$$

Random SVD Factorization

The problem of computing a low-rank approximation to a given matrix can be divided into two stages: construction of a low-dimensional subspace and computation of the factorization with the restriction to the subspace.

Algorithm 2 Soft-Impute by Nesterov's Acceleration

Generate a random $n \times k$ matrix R
 Set the precision $\epsilon = 10^4$
 $Q = \text{powermethod}(Y, R, \epsilon)$
 $[U', D', V'] = \text{svd}(Q^T Z)$
 $U = U', \quad D = D', \quad V = V'$
 $c = c' = 1$
for (λ_i, γ_i) with the descending sequences $\{\lambda_i\}_{i=1}^t$ and $\{\gamma_i\}_{i=1}^t$ **do**
 $\theta = \frac{c'-1}{c}$
 $X^* = (1 + \theta) \cdot UDV^T - \theta \cdot U'D'V'^T$
 $Y^* = P_{\Omega}(Y - X^*) + X^*$
 $R = \text{qr}(V)$
 $U' = U, \quad V' = V, \quad D' = D$
 $Q = \text{powermethod}(Y^*, R, \epsilon)$
 $[U, \text{diag}(\sigma), V] = \text{svd}(Q^T Y^*)$
 $U = QU$
 $D = \text{diag}(s_{\lambda_i, \gamma_i}(\sigma))$
 $c' = c$
 $c = \frac{1 + \sqrt{1 + 4c'^2}}{2}$
end for

Algorithm 3 Power method (powermethod) for computing orthonormal subspace

Given Z, R and ϵ
 $Z' = ZR$
 $Q = \text{qr}(Z')$
for $i = 1, 2, \dots, q$ **do**
 $Z' = ZZ^T Q$
 $Q' = \text{qr}(Z')$
 $\text{error} = \|Q' - Q\|_F^2$
 $Q = Q'$
 if $\text{error} < \epsilon$ **then**
 break
 end if
end for
 return Q

We require a matrix Q here as an approximate basis (see Alg. 3) for the range of the input matrix Z

$$Z \approx QQ^T Z, \tag{24}$$

where Q has orthonormal columns.

In the simplest form, we directly compute the QR factorization of $Y = Z\Omega = QR$, where Ω is a random matrix. Thus, Q is an orthonormal base of the range of Z . However, the singular spectrum of the input matrix may decay slowly. Therefore we alternatively

compute the QR factorization of $(ZZ^T)^q Z\Omega = QR$, where $q \leq 5$ usually suffices in practice.

Given the orthonormal matrix Q we compute the SVD of $B = Q^T Z$ as follows:

$$B = Q^T Z = \tilde{U} \Sigma V^T, \quad (25)$$

then with (24), we immediately obtain the approximate SVD of Z :

$$Z \approx QQ^T Z = QB = (Q\tilde{U}) \Sigma V^T. \quad (26)$$

Endnote

¹<https://grouplens.org/datasets/movielens/>.

Abbreviations

ALM: Augmented Lagrangian method; ALS: Alternative least squares; FPCA: Fixed point continuation with approximate SVD; MMMF: Maximum-margin matrix factorization; NMCNA: Nonconvex matrix Completion by Nesterov's acceleration; RMSE: Root mean squared error SDP: Semidefinite programming; SVT: Singular value thresholding

Acknowledgements

The authors acknowledge supports from their affiliations.

Funding

This work was partially supported by the Fundamental Research Funds for the Henan Provincial Colleges and Universities in the Henan University of Technology (2016RCJH06), Jiangsu University Natural Science Programme Grant 17KJB520041, the National Natural Science Foundation of China (61103138 and 61473236).

Availability of data and materials

The dataset are provided by MovieLens can be downloaded in the website <https://grouplens.org/datasets/movielens/>.

Authors' contributions

Conceived the research: J-XB. Designed and Performed the experiments: J-XB and X-GS. Analyzed the data: W-QF and Z-GQ. Contributed materials/analysis tools: Z-GQ. Wrote the paper: J-XB, Z-GQ and G-GG. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Henan University of Technology, Lianhua, Zhengzhou, China. ²Henan University of Science and Technology, Luoyang, China. ³Xian Jiaotong-Liverpool University, Suzhou, China. ⁴Ocean University of China, Qingdao, China. ⁵China Internet Network Information Center, Beijing, China.

Received: 5 September 2018 Accepted: 26 November 2018

Published online: 20 December 2018

References

- Huang K-Z, Yang H, King I, Lyu MR. Machine Learning: Modeling Data Locally And Globally. In: Advanced Topics in Science and Technology in China. Berlin Heidelberg: Springer; 2008.
- Jin X-B, Geng G-G, Xie G-S, Huang K. Approximately optimizing NDCG using pair-wise loss. *Inf Sci.* 2018;453:50–65.
- Jin X-B, Geng G-G, Sun M, Zhang D. Combination of multiple bipartite ranking for multipartite web content quality evaluation. *Neurocomputing.* 2015;149:1305–14.
- Jin X-B, Xie G-S, Huang K, Hussain A. Accelerating Infinite Ensemble of Clustering by Pivot Features. *Cogn Comput.* 2018;6:1–9.

5. Yang X, Huang K, Zhang R, Hussain A. Learning Latent Features With Infinite Nonnegative Binary Matrix Trifactorization. In: Transactions on Emerging Topics in Computational Intelligence, IEEE; 2018. p. 1–14.
6. Candès E., Recht B. Exact Matrix Completion via Convex Optimization. *Commun ACM*. 2012;55(6):111–9.
7. Cai J-F, Candès EJ, Shen Z. A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM J on Optimization*. 2010;20(4):1956–82. <https://doi.org/10.1137/080738970>.
8. Lin Z, Chen M, Ma Y. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *J Struct Biol*. 2013;181(2):116–27. arXiv: 1009.5055.
9. Ma S, Goldfarb D, Chen L. Fixed point and Bregman iterative methods for matrix rank minimization. *Math Program*. 2011;128(1–2):321–53.
10. Shen Z, Toh K-c, Yun S. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific J Optim*. 6:615–640.
11. Tomioka R, Suzuki T, Sugiyama M, Kashima H. A Fast Augmented Lagrangian Algorithm for Learning Low-rank Matrices. In: ICML. ICML'10. USA: Omnipress; 2010. p. 1087–94.
12. Mazumder R, Friedman JH, Hastie T. SparseNet: Coordinate Descent With Nonconvex Penalties. *J Am Stat Assoc*. 2011;106(495):1125–38.
13. Bertsimas D, King A, Mazumder R. Best Subset Selection via a Modern Optimization Lens; 2015. arXiv:1507.03133 [math, stat]. arXiv: 1507.03133.
14. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat*. 2010;38(2):894–942. arXiv: 1002.4734.
15. Zhang C-H, Zhang T. A General Theory of Concave Regularization for High-Dimensional Sparse Estimation Problems. *Stat Sci*. 2012;27(4):576–93.
16. Mazumder R, Saldana DF, Weng H. Matrix Completion with Nonconvex Regularization: Spectral Operators and Scalable Algorithms; 2018. arXiv:1801.08227 [stat]. arXiv: 1801.08227.
17. Halko N, Martinsson P-G, Tropp JA. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions; 2009. arXiv:0909.4061 [math]. arXiv: 0909.4061.
18. Nesterov Y. Gradient methods for minimizing composite functions. *Math Program*. 2013;140(1):125–61.
19. Sutskever I, Martens J, Dahl G, Hinton G. On the Importance of Initialization and Momentum in Deep Learning. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. ICML'13. Atlanta: JMLR.org; 2013. p. 1139–47.
20. Gu R, Dogandzic A. Projected Nesterov's Proximal-Gradient Algorithm for Sparse Signal Recovery. *IEEE Trans Sig Process*. 2017;65(13):3510–25.
21. Hastie T, Mazumder R, Lee J, Zadeh R. Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares; 2014. arXiv:1410.2596 [stat]. arXiv: 1410.2596.
22. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. In: Springer Series in Statistics 2nd edn. New York: Springer; 2009.
23. Golub GH, Van Loan CF. *Matrix Computations* (3rd Ed.). Baltimore: Johns Hopkins University Press; 1996.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

