2016 IEEE 16th International Conference on Data Mining

Aligned Matrix Completion: Integrating Consistency and Independency in Multiple Domains

Linli Xu*, Zaiyi Chen*, Qi Zhou*, Enhong Chen*, Nicholas Jing Yuan†, Xing Xie†
*School of Computer Science and Technology, University of Science and Technology of China
Email: linlixu@ustc.edu.cn, czy6516@mail.ustc.edu.cn, zhouqixs@mail.ustc.edu.cn, cheneh@ustc.edu.cn
†Microsoft Research, Email: nicholas.yuan@microsoft.com, xing.xie@microsoft.com

***Abstract*—Matrix completion is the task of recovering a data matrix from a sample of entries, and has received significant attention in theory and practice. Normally, matrix completion considers a single matrix, which can be a noisy image or a rating matrix in recommendation. In practice however, data is often obtained from multiple domains rather than a single domain. For example, in recommendation, multiple matrices may exist as *user*×*movie* and *user*×*book*, while correlations among the multiple domains can be reasonably exploited to improve the quality of matrix completion. In this paper, we consider the problem of *aligned matrix completion*, where multiple matrices are recovered that correspond to different representations of the same group of objects. In the proposed model, we maintain consistency of multiple domains with a shared latent structure, while allowing independent patterns for each separate domain. In addition, we impose the low-rank structure of a matrix with a novel regularizer which provides better approximation than the standard nuclear norm relaxation.**

I. Introduction

Matrix completion is a widely investigated problem with significant theoretical and practical interests where one intends to recover a data matrix from a small fraction of observed entries. Under certain conditions, i.e., the partially observed matrix is low-rank and incoherent, various algorithms can be designed to reconstruct the matrix [1]. The technique of matrix completion has been successfully applied to various tasks including collaborative filtering [2], video denoising [3], transductive learning [4], etc., where a single matrix is considered and reconstructed.

In practice however, data is often obtained from multiple domains rather than a single domain. For example, in computer vision, an object can be captured by cameras from different angles; in recommendation, a user can rate items in different domains of movies, books, or music. In this context, information from multiple domains can be represented as multiple aligned matrices, which correspond to different representations of the same group of objects.

In general, there exist intrinsic correlations among the multiple domains. For example, a user that rates "romance" higher than "horror" in the *movie* domain may have the same preference in the *book* domain. Intuitively, the correlations, if appropriately exploited, can be helpful to model the objects better and improve the quality of prediction. This motivates the multi-view learning principle [5], [6] that exploits the underlying consistency among different views. However, multi-view learning normally considers complete data from multiple sources; while in many circumstances the data matrices from multiple domains may be incomplete and need to be reconstructed. In the meantime, the technique of collective matrix factorization (CMF) [7] learns low-rank representations given a collection of matrices with shared factors, and can work on the task of reconstructing multiple related matrices. Nevertheless, the principle of CMF is restricted in the sense that it assumes all the domains share the same latent representations; while in practice, there exist scenarios where individual matrices may have strong domain-specific patterns, and lack of distinction between the consistent and domain-specific factors may imply improper transfer of information among different domains and degeneration of prediction performance.

On the other hand, in matrix completion, it is common to assume that the partially observed matrix is low-rank, which can be enforced with various rank regularizers. Given the NP-hardness of the rank minimization problems, a widely used relaxation of the rank function is the nuclear norm. It is shown in [1] that low-rank solutions can be recovered perfectly via the nuclear norm under incoherence assumptions. Unfortunately, in real applications, the underlying matrix may have no incoherence property and the data may be grossly corrupted. Moreover, the nuclear norm suffers from the limitation that it adds up all the singular values with equal weights which implies that large singular values are penalized more than small ones; whereas the large singular values, corresponding to more important components, should be penalized less to preserve the major information. This issue can get even worse when reconstructing multiple matrices simultaneously with varied degrees of sparsity.

In this paper, we propose the novel model of Aligned Matrix Completion (Aligned MC), where multiple matrices are recovered simultaneously that correspond to different views of the same group of objects. The above two issues are addressed in the proposed model. We factorize the latent representations for multiple domains by maintaining consistency with a shared latent structure while allowing independent factors for each separate domain. In addition, to overcome the imbalanced penalization of different singular values, we impose the low-rank structure of a matrix with a general singular value regularization and further extend it to the scenario of multiple domains. Theoretical analysis is then included with conver-

2374-8486/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDM.2016.112

gence guarantees. The proposed framework is evaluated on synthetic data as well as the empirical task of recommendation in multiple domains which predicts a user's preference on multiple types of items.

The rest of the paper is organized as follows. In Section II related work is discussed. Next in Section III we propose the novel method of Aligned Matrix Completion (Aligned MC). Rigorous convergence analysis of the proposed algorithm is conducted in Section IV, followed with experimental results summarized in Section V to demonstrate the empirical effectiveness of the proposed method. The paper is then concluded in Section VI.

II. Related Work

In matrix completion, it is common to assume that a partially observed matrix has low rank structure, which entails a rank minimization problem. To tackle this problem which is NP-hard in general, the nuclear norm is normally used as a convex relaxation of the rank function. To overcome the issue of imbalanced penalization of different singular values of the nuclear norm, various non-convex rank relaxations are proposed with a weighted sum of singular values while choosing appropriate and fixed weights in a non-descending order [8]-[10]. A more recent work [11] tries to minimize a reweighted nuclear norm for a better approximation of the rank function as well as the observed matrix with provable convergence. In this paper, we present a family of singular value regularization functions and generalize a group of non-convex rank relaxations with a more elaborate convergence analysis. We further extend the methodology to aligned matrix completion in multiple domains, and tackle the empirical task of recommendation in multiple domains.

Various efforts have been devoted to recommendation in multiple domains, among which transfer learning is a widely applied principle [12]-[14], where the model for each domain needs to be trained separately with the source domain and the target domain specified. On the other hand, multiple recommendation tasks on different domains can be performed simultaneously by effectively exploiting the correlations between domains [7], [15]. Specifically, collective matrix factorization (CMF) [7] jointly factorizes multiple matrices assuming common latent factors for all the domains, which may be hardly true in practice especially in scenarios with strong domain-specific patterns for each domain. To address that, a recent work of group-sparse matrix factorization (GSMF) [16] incorporates group sparsity on the latent factors, allowing different factors selected for different domains. However it does not necessarily entail a common subset of factors for each domain. In this paper, we consider the task of reconstructing matrices in multiple domains simultaneously. To achieve that we factorize the latent representations of multiple domains with a shared latent structure and independent factors for each separate domain to integrate consistency and independency across various domains in the model.

III. Aligned Matrix Completion

A. Generalized Low-rank Matrix Completion

We first consider the low-rank matrix completion problem in a single domain. Given a noisy matrix $Y \in \mathbb{R}^{n\times m}$ with N observations, the principle of low-rank matrix completion tries to find a matrix X that the entries of X indexed by $\Omega = \{(i,j)|X_{ij} \text{ is observed}\}$ are as close to Y as possible, namely $X_\Omega \approx Y_\Omega$, and $\text{rank}(X) \leq l$. Considering the risk of estimating X_Ω with a loss function ℓ, the low-rank matrix completion problem can be formulated as

$$\min_X \ell(X_\Omega) + \lambda \cdot \text{rank}(X). \tag{1}$$

This rank minimization problem is NP-hard in general due to the non-convexity and discontinuity of the rank function. A common strategy is to relax it with the nuclear norm $||\cdot||_*$ as a low-rank approximation:

$$\min_X \ell(X_\Omega) + \lambda||X||_* \tag{2}$$

Although the low-rank approximation (2) is the tightest convex relaxation of (1) [17], the nuclear norm may not be a good approximation of the rank function due to the fact that it adds up all the singular values equally, which implies that large singular values are penalized more heavily than small ones. Therefore, we propose a family of Generalized Singular Value Regularization (GSVR) functions

$$h(X) = \sum_i r_i(\sigma_i(X)) \tag{3}$$

where $\sigma_i(X)$ denotes the i-th largest singular value of X and each r_i is a general function of the corresponding singular value $\sigma_i(X)$, which can be flexibly designed to reflect the inherent structure of the matrix.

Essentially, GSVR represents a family of singular value regularization functions and generalizes a group of methods including Truncated Nuclear Norm Regularization (TNNR) [8], Reweighted Nuclear Norm (RNN) [11], etc. Note that $h(X)$ is allowed to be convex or non-convex regarding X. Nevertheless, we will design a proximal algorithm with convergence guarantees to solve the matrix completion problem with the generalized singular value regularization.

B. Aligned Matrix Completion in Multiple Domains

In multi-domain scenarios, given observations indexed by $\{\Omega_d, d = 1,\ldots,D\}$ from D domains: $\{Y^d \in \mathbb{R}^{n\times m_d},\ d = 1,\ldots,D\}$ where matrices $\{Y^d\}$ are aligned in rows, correlations among the multiple domains can be exploited to improve the quality of matrix completion. Specifically, we assume there exist consistency shared among multiple domains as well as independent patterns for each separate domain. In the case of multi-domain recommendation where matrices $\{Y^d\}$ correspond to rating matrices on different types of items such as *user*×*movie* and *user*×*book*, it is natural to assume that *users* have some mutual interests across domains, as well as some distinct interests in each domain.

Consider the latent factors of *users* and *items* by factorizing a rating matrix $X = UV^\top$, where U and V correspond to low-rank *user×latent factor* and *item×latent factor* matrices. In multiple domains, the consistent patterns can be represented by a shared *user×latent factor* matrix U. As a consequence, the observations in the d-th domain can be factorized as $Y^d_{\Omega_d} = (UV^{d^\top} + \tilde{U}^d\tilde{V}^{d^\top} + \varepsilon^d)_{\Omega_d}$, where $X^d = UV^{d^\top}$ represents shared user interests on the d-th domain; and $\tilde{X}^d = \tilde{U}^d\tilde{V}^{d^\top}$ corresponds to domain specific user preference. The rating behaviors of shared user interests on various domains can be summarized in the matrix $X = [X^1, \ldots, X^D] = U \cdot [V^{1^\top}, \ldots, V^{D^\top}]$, which is a horizontal concatenation of $\{X^d\}$. To learn the shared and domain specific user interests, we apply a general singular value regularizer h_0 on X, and h_d on X^d for $d = 1, \ldots, D$. The optimization problem can then be formulated as

$$\min_{\substack{X,\,\{\tilde{X}^d\},\\ d\,=\,1,\ldots,\,D}} \ell(X, \tilde{X}^1, \ldots, \tilde{X}^D) + h_0(X) + \sum_{d=1}^{D} h_d(\tilde{X}^d). \quad (4)$$

In this paper, we use $\ell = \sum_{d=1}^{D} \frac{N_{\max}}{N_d} ||Y^d_{\Omega_d} - (X^d + \tilde{X}^d)_{\Omega_d}||_F^2$ to measure the reconstruction error and balance the losses of different domains, where N_d is the number of observations in domain d and $N_{\max}$ is the maximum of $\{N_d\}$. Regularization functions h_0, $\{h_d\}$ are designed as the weighted sum of the singular values with non-descending weights w_i, which is a special form of (3):

$$\begin{aligned} h(X) &= \sum_{i=1}^{\mathrm{rank}(X)} w_i \sigma_i(X), \\ w_i &= \mathrm{pen}_1 + \frac{\mathrm{pen}_2}{1 + e^{-\gamma(i-k)}}. \end{aligned} \quad (5)$$

Here pen_1 and pen_2 are positive constants. w_i is designed such that the first k singular values are penalized less to preserve the major information of a matrix, where γ determines the sharpness of the sigmoid function. In the meantime, the non-descending singular value regularization $h(X)$ is no longer convex with respect to X. Nevertheless, a sub-linear convergence rate can still be achieved by our algorithm as proved in the following section.

IV. Optimization and Convergence Analysis

In this section, we will build the optimization algorithm and discuss the convergence properties.

A. Proximal Gradient Algorithms for Single and Multiple Variables

The matrix completion problem with generalized singular value regularization (GSVR) (3) for single and multiple variables can be formulated as

$$\min_{X \in \mathbb{R}^{n\times m}} \Phi(X) = \ell(X) + h(X), \quad (6)$$

and

$$\min_{X^1,\ldots,X^D} \Phi(X^1, ..., X^D) = \ell(X^1, \ldots, X^D) + \sum_{d=1}^{D} h_d(X^d), \quad (7)$$

respectively. With a slight abuse of notation, D here represents the number of matrix variables, rather than the number of domains in the previous section.

To solve the problems, we first define the proximal map $P_h^\mu,\ \mu > 0$ for h:

$$P_h^\mu(M_t) = \arg\min_{X \in \mathbb{R}^{n\times m}} \frac{1}{2}||X - M_t||^2 + \mu h(X), \quad (8)$$

which is actually solving the problem

$$\begin{aligned} X_{t+1} = \arg\min_{X \in \mathbb{R}^{n\times m}} \ell(X_t) + \langle \nabla\ell(X_t), X - X_t\rangle \\ + \frac{1}{2\mu}||X - X_t||^2 + h(X), \end{aligned} \quad (9)$$

where $M_t = X_t - \mu\nabla\ell(X_t)$.

Problem (6) can now be solved by iteratively solving for $P_h^\mu(M_t)$ to find the closest point X_{t+1} in the feasible set defined by $h(M_t)$, after going along the direction of $-\nabla\ell(X_t)$ with a small step to get an intermediate solution M_t. However, each r_i is not separable with respect to X. Based on the following lemma, a solution of (8) with penalties in (3) can be found in an easier way.

Lemma 1. *[18] Let $||\cdot||$ be a unitarily invariant norm on $\mathbb{R}^{n\times m}$ (i.e., $||LXR|| = ||X||$ for any unitary matrix L, R) and let $F : \mathbb{R}^{n\times m} \to \mathbb{R}$ be a unitarily invariant function (i.e., $F(LXR) = F(X)$ for any unitary matrix L, R and any $X \in \mathbb{R}^{n\times m}$). Let $A = U\Sigma V^\top \in \mathbb{R}^{n\times m}$ be given,* $\mathrm{Diag}(\mathbf{x})$ *be a diagonal matrix with $\mathbf{x}$ on its diagonal, and h be a non-decreasing function on $[0, \infty)$. Then $X^* = U\mathrm{Diag}(\mathbf{x}^*)V^\top$ is a global optimal solution of the problem*

$$\min_X F(X) + h(||X - A||) \quad (10)$$

where $\mathbf{x}^*$ is the global optimal solution of the problem

$$\min_{\mathbf{x}} F(\mathrm{Diag}(\mathbf{x})) + h(||\mathrm{Diag}(\mathbf{x}) - \Sigma||). \quad (11)$$

It is worthwhile to further our discussion regarding this lemma. If we set $g(\mathbf{x}) = F(\mathrm{diag}(\mathbf{x}))$, then g can be viewed as an extension of the *symmetric gauge function* for F, in which case g is a function on $\mathbb{R}^n$ whose value is invariant under permutations but could be variant under sign changes of components. Due to these facts, we can view a unitarily invariant function F as an extension of a unitarily invariant norm. More examples of symmetric gauge functions in normed vector space and analyses can be found in [19]. As a result, if the empirical risk ℓ is measured by a norm in vector space, or more generally by a unitarily invariant function, and non-smooth regularization terms $\{h_d\}$ penalize the unitarily invariant norms of variables non-decreasingly, Lemma 1 indicates that the proximal map could be computed in an easier way.

Algorithm 1 GSVR Proximal Gradient (GSVR-PG) Algorithm for a Single Variable

Input: Observed matrix Y for each view, Lipschitz constant L and stop criterion ε.
Output: Recovered matrix X
Initialize: $X = 0$, $\mu < 1/L$
while $||X_{t+1} - X_t||^2 \geq \varepsilon||Y||^2$ **do**
$M_t = X_t - \mu\nabla_{X_t}\ell(X_t)$
$X_{t+1} = P_h^{\mu}(M_t)$
end while

Algorithm 2 GSVR Proximal Gradient (GSVR-PG) Algorithm for Multiple Variables

Input: Observed matrices $\{Y^d\}$ for each view, the largest Lipschitz constant $L_{\max}$ and stop criterion ε.
Output: Recovered matrices $\{X^d\}$
Initialize: $X^d = 0$, $\mu < 1/L_{\max}$
while $\exists X^d$ such that $||X_{t+1}^d - X_t^d||^2 \geq \varepsilon||Y^d||^2$ **do**
for $d = 1, \ldots, D$ **do**
$M_t^d = X_t^d - \mu\nabla_{X_t^d}\ell(X_t^d)$
$X_{t+1}^d = P_{h_d}^{\mu}(M_t^d)$
end for
end while

Corollary 1. *The proximal map P_h^{μ} of h in the form of (5) can be computed as:*

$$P_h^{\mu}(M_t) = U\mathrm{Diag}(\mathbf{x}^*)V^{\top},$$
$$x_i = \begin{cases} \sigma_i(M_t) - \mu w_i, & \text{if } \sigma_i(M_t) > w_i \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Proof. It is obvious that the Frobenius norm is unitarily invariant, $h(\theta) = \theta^2$ is nondecreasing on $[0, \infty)$, and penalties defined as in (3) are also unitarily invariant and separable for each singular value. Given that all assumptions of Lemma 1 are satisfied, the proximal maps of $\{r_i\}$ can be calculated by

$$P_{r_i}^{\mu}(M_t) = U\mathrm{Diag}([0, \ldots, x_i^*, \ldots, 0])V^{\top},$$
$$x_i^* = \arg\min_{x_i \in \mathbb{R}} \frac{1}{2}||\sigma_i(M_t) - x_i||^2 + \mu\, r_i(x_i). \tag{13}$$

The second equation of (13) is a univariate optimization problem, which is much easier to solve.

Based on the above, the proximal map P_h^{μ} can be computed separately as

$$P_h^{\mu}(M_t) = \sum_i \alpha_i P_{r_i}^{\mu}(M_t), \tag{14}$$

which means P_h^{μ} is strictly equal to the convex combination of $\{P_{r_i}^{\mu}\}$. Substituting (12) into (13) and (14), we can complete the proof. □

The proximal method for a single matrix variable is described in Algorithm 1.

In the multivariate scenario, we use an alternating update scheme which updates each variable with a small step in sequence:

$$X_{t+1}^d = \arg\min_{X^d \in \mathbb{R}^{n \times m}} \ell(X_t^d) + \langle \nabla\ell(X_t^d), X^d - X_t^d \rangle + \frac{1}{2\mu}||X^d - X_t^d||^2 + h_d(X^d), \tag{15}$$
$$d = 1, \ldots, D \text{ sequentially.}$$

The algorithm designed for this update strategy is summarized in Algorithm 2.

B. Convergence Analyses

In this subsection, we will analyze the convergence of sequences generated by Algorithm 1 and Algorithm 2 for the single- and multiple-matrix completion problems respectively. It is worth noting that the mild conditions required in the proofs, including Assumptions **(A1), (A2), (A3)** and **(A4)**, are satisfied by a large number of functions, which will not affect the generalization ability of the proximal algorithm in general. Moreover, compared to existing work [8], [10], the more explicit analyses established in this subsection will guarantee that a large number of classical objective functions can be optimized by our proposed algorithm, which will converge to a critical point with a superior upper bound of the number of required iterations.

This subsection is arranged as follows. We start from the assumptions of ℓ and h, and explain the KL property which plays an important role in proving the convergence. Then the convergence guarantee in the scenario of single matrix variable is derived. Next, we will go though the proof of convergence in the multivariate scenario, followed by showing the effectiveness of the alternating proximal method for our specific problem of aligned matrix completion (4) explicitly.

Before going through the details, we make the following assumptions about ℓ and h to facilitate the analysis.

(A1) Function $\ell : \mathbb{R}^{n \times m} \to \mathbb{R}$ is lower bounded, continuously differentiable with *L-Lipschitz* continuous gradient (*w.r.t. the Euclidean distance, or Frobenius norm for matrices*). That is, there exists a positive constant L such that

$$||\nabla\ell(A) - \nabla\ell(B)|| \leq L||A - B||, \forall A, B \in \mathbf{dom}\ell \tag{16}$$

(A2) Each penalty component $r_i : \mathbb{R} \to \mathbb{R}$ is a proper, lower bounded function.

(A3) Function Φ has the KL property.

As an important property in the following analysis, the definition of the Kurdyka-Łojasiewicz (KL) property [20] is summarized below. Before that we first define the distance from any subset $S \subset \mathbb{R}^n$ to any point $x \in \mathbb{R}^n$ as

$$\mathrm{dist}(x, S) = \inf\{||y - x||, y \in S\}. \tag{17}$$

Definition 1. (KL property) *Let $\sigma : \mathbb{R}^n \to (-\infty, +\infty]$ be proper and lower semi-continuous.*

(i) A function σ has the KL property *at $\bar{\mu} \in \operatorname{dom} \partial\sigma := \{u \in \mathbb{R}^n : \partial\sigma(u) \neq \varnothing\}$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of $\bar{u}$ and a function $\psi \in \Psi_\eta$, such that for all*

$$u \in U \cap [\sigma(\bar{u}) < \sigma(u) < \sigma(\bar{u}) + \eta], \tag{18}$$

the following inequality holds

$$\psi'(\sigma(u) - \sigma(\bar{u}))\operatorname{dist}(0, \partial\sigma(u)) \geq 1. \tag{19}$$

(ii) If σ satisfies the KL property *at each point of* $\operatorname{dom} \partial\sigma$*, then σ is called a* KL function*.*

Now we will go further to show how the KL property works in the proof of convergence.

Lemma 2. (Uniformized KL property) [21] *Let Θ be a compact set and let $\sigma : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper and lower semi-continuous function. Assume that σ is constant on Θ and satisfies the KL property at each point of Θ. Then there exist $\varepsilon > 0, \eta > 0$ and $\phi \in \Phi_\eta$ such that for all $\bar{u}$ in Θ and all u in the following intersection:*

$$\{u \in \mathbb{R}^n : \operatorname{dist}(u, \Theta) < \varepsilon\} \cap [\sigma(\hat{u}) < \sigma(u) < \sigma(\hat{u}) + \eta], \tag{20}$$

the following inequality holds

$$\phi'(\sigma(u) - \sigma(\hat{u}))\ \operatorname{dist}(0, \partial\sigma(u)) \geq 1. \tag{21}$$

This lemma indicates that, if the KL property holds in the neighborhood of critical points [21], the proximal algorithm is guaranteed to converge to a critical point in finite steps. As a consequence, the convergence analysis will be considerably simplified by using this property.

Equipped with this tool, one can prove that Algorithm 1 for a single matrix variable will converge to a critical point after entering its neighborhood. The overall convergence properties can be summarized in the following theorem.

Theorem 1. *If assumptions **(A1)**, **(A2)** and **(A3)** hold, penalty h is defined as in (3), and functions ℓ and $\{r_i\}$ are definable; given a step size $\mu < 1/L$, the sequence $\{X_t\}_{t\in\mathbb{N}}$ generated by the Generalized Singular Value Regularization-Proximal Gradient (GSVR-PG) algorithm has finite length and converges to a critical point of (6). That is*

(i) The sequence $\{X_t\}_{t\in\mathbb{N}}$ has finite length,

$$\sum_{t=1}^{\infty} ||X_{t+1} - X_t|| < \infty \tag{22}$$

(ii) The sequence $\{X_t\}_{t\in\mathbb{N}}$ converges to a critical point X^* of (6).

Following Lemma 1 and Assumptions **(A1)** **(A2)** **(A3)**, we can derive this theorem according to [22] and [23]. It is worth noting that, if r_i is a concave function, one can use its first order approximation to bound the proximal map from above as demonstrated in [24] and [11].

Next we will prove the convergence of Algorithm 2 for problem (4) with the multivariate function $\ell(\cdot, \ldots, \cdot) : \mathbb{R}^{n_1 \times m_1} \times \ldots \times \mathbb{R}^{n_D \times m_D} \to \mathbb{R}$. To begin we also need an assumption similar to **(A1)** regarding its structure.

(A4) Multivariate function $\ell(X^1, \ldots, X^D)$ is lower bounded, continuously differentiable, and has L_d-*Lipschitz* continuous partial gradient with respect to each X^d. Meanwhile, $\nabla\ell$ is Lipschitz continuous on bounded subsets of $\mathbb{R}^{n_1 \times m_1} \times \ldots \times \mathbb{R}^{n_D \times m_D} \to \mathbb{R}$. That is, for each bounded subsets $B_1 \times \ldots \times B_D$, there exists a constant $M > 0$, such that for all $(X^1, \ldots, X^D) \in B_1 \times \ldots \times B_D$, the following inequality holds:

$$\begin{aligned} &||(\nabla_{X^1}\ell(X_1^1, \ldots, X_1^D) - \nabla_{X^1}\ell(X_2^1, \ldots, X_2^D), \ldots, \\ &\quad \nabla_{X^D}\ell(X_1^1, \ldots, X_2^D) - \nabla_{X^D}\ell(X_2^1, \ldots, X_2^D))|| \\ &\leq M||(X_1^1 - X_2^1, \ldots, X_1^D - X_2^D)||. \end{aligned} \tag{23}$$

Here the main difference with the single variable case is that we make one more assumption on the gradient. We can see that ℓ in our problem (4) is C^2 continuous and following the Mean Value Theorem, Assumption **(A4)** will be satisfied.

To analyse the convergence property of Algorithm 2, we first show that the sequence generated by Algorithm 2 would converge to some limit points if assumptions hold, and these limit points would be a subset of critical points of Φ. This result is also known as subsequence convergence. Then, based on the properties of the KL function, we can guarantee that the algorithm will converge to one of the critical points, which is also known as the global convergence.

For simplicity, we use the following abbreviations in the $(t+1)$-th iteration:

$$\begin{aligned} \ell_{t+1}(X_t^d) &= \ell(X_{t+1}^1, \ldots, X_{t+1}^{d-1}, X_t^d, \ldots, X_t^D), \\ \ell_{t+1}(X_{t+1}^d) &= \ell(X_{t+1}^1, \ldots, X_{t+1}^d, X_t^{d+1}, \ldots, X_t^D). \end{aligned} \tag{24}$$

We also define

$$\rho = \min\{\mu^{-1} - L_1, \ldots, \mu^{-1} - L_D\}, \tag{25}$$

the sequence generated by Algorithm 2 as

$$Z_t = (X_t^1, \ldots, X_t^D),\ \forall t \geq 0, \tag{26}$$

and

$$\sum_{d=1}^{D} ||X_{t-1}^d - X_t^d||^2 = ||Z_{t-1} - Z_t||^2. \tag{27}$$

Then following (24), we get

$$\Phi_t(Z_t) = \ell_t(Z_t) + \sum_{d=1}^{D} h_d(X_t^d). \tag{28}$$

To prove the global convergence, we start with extending the proof of convergence properties from single-variate case to multivariate case, which are summarized in Lemma 3 and Lemma 4.

Lemma 3. (Convergence properties) *Suppose that Assumptions **(A2)** and **(A4)** hold. The following assertions hold.*

(i) The sequence $\{\Phi(Z_t)\}_{t\in\mathbb{N}}$ is non-increasing and

$$\frac{\rho}{2}||Z_{t+1} - Z_t||^2 \leq \Phi(Z_t) - \Phi(Z_{t+1}),\ \forall t \geq 0. \tag{29}$$

(ii) We have

$$\sum_{t=1}^{\infty}\sum_{d=1}^{D}||X_{t+1}^d - X_t^d||^2 = \sum_{t=1}^{\infty}||Z_{t+1} - Z_t||^2 < \infty, \quad (30)$$

then $\lim_{t\to\infty}||Z_{t+1} - Z_t|| = 0$.

Proof. Since $X_{t+1}^d, d = 1, \ldots, D$, is the optimal solution of problem (9), in the $(t+1)$-th iteration we have

$$\begin{aligned}&\langle \nabla_{X_t^d}\ell_{t+1}(X_t^d), X_{t+1}^d - X_t^d\rangle + h_d(X_{t+1}^d)\\ &+ \frac{1}{2\mu}||X_{t+1}^d - X_t^d||^2 \le h_d(X_t^d)\end{aligned} \quad (31)$$

Following assumption **(A4)**, we have

$$\begin{aligned}\ell_{t+1}(X_{t+1}^d) \le \ell_{t+1}(X_t^d) + \tfrac{L_d}{2}||X_{t+1}^d - X_t^d||^2\\ +\langle \nabla_{X_t^d}\ell_{t+1}(X_t^d), X_{t+1}^d - X_t^d\rangle\end{aligned} \quad (32)$$

Combining (31), (32) we get

$$\begin{aligned}\ell_{t+1}(X_{t+1}^d) + h_d(X_{t+1}^d) \le \ell_{t+1}(X_t^d) + h_d(X_t^d)\\ - \frac{\mu^{-1} - L_d}{2}||X_{t+1}^d - X_t^d||^2\end{aligned} \quad (33)$$

Adding up the above inequalities regarding $d = 1, 2, ..., D$, for all $t \ge 0$ we have

$$\begin{aligned}\Phi(Z_t) - \Phi(Z_{t+1}) =& \sum_{d=1}^{D}[\ell_t(X_t^d) + h_d(X_t^d)\\ &- \ell_{t+1}(X_{t+1}^d) - h_d(X_{t+1}^d)]\\ \ge& \sum_{d=1}^{D}\frac{\mu^{-1} - L_d}{2}||X_{t+1} - X_t||^2.\end{aligned} \quad (34)$$

Following (34), we have that the sequence $\{\Phi(Z_t)\}_{t\in\mathbb{N}}$ is non-increasing, and since Φ is bounded from blow according to Assumption **(A4)**, it will converge to some real number $\overline{\phi}$. Meanwhile, Since we choose the step size smaller than the reciprocal of the largest Lipschitz constant $L_{\max}$ as shown in Algorithm 2, from (25) it follows that

$$\sum_{d=1}^{D}\frac{\mu^{-1} - L_d}{2}||X_{t+1} - X_t||^2 \ge \frac{\rho}{2}||Z_{t+1} - Z_t||^2. \quad (35)$$

Combining (34) and (35), (i) is proved.

By summing up (29) from $t = 0$ to $N - 1$ and taking the limit $N \to \infty$, we can prove (ii). □

Based on Lemma 3, we can conclude that in $O(1/\varepsilon)$ iterations, Algorithm 2 will stop. This assertion is summarized as follows.

Corollary 2. *Let $\{(X_t^1, \ldots, X_t^D)\}$ be the sequence generated by Algorithm 2 with $\mu < 1/L_{\max}$, which converges to some limit points $\{(X^{1^*}, \ldots, X^{D^*})\}$. Then for all $T \ge 0$, we have*

$$\begin{aligned}&\min_{0\le t\le T}\sum_{d=1,\ldots,D}||X_{t+1}^d - X_t^d||^2\\ &\le \frac{2(\Phi(X_0^1, \ldots, X_0^D) - \Phi(X^{1^*}, \ldots, X^{D^*}))}{\rho T}.\end{aligned} \quad (36)$$

This corollary can be achieved by summing up (34) and rearranging the inequality.

Next, to understand the characteristics of the points that Algorithm 2 will converge to, we need the following lemma to analyze the limit point(s).

Lemma 4. (The lower bound of the iterate gap based on subgradient) *Suppose that assumptions **(A2)** and **(A4)** hold. Let $\{z^k\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 2 which is assumed to be bounded. For each iteration $t > 0$ and $d = 1, ..., D$, define*

$$\begin{aligned}A_t^d =& \mu^{-1}(X_{(t-1)}^d - X_t^d) + \nabla_{X^d}\ell_t(Z_t)\\ &- \nabla_{X^d}\ell_t(X_{(t-1)}^d),\\ &d = 1, \ldots, D.\end{aligned} \quad (37)$$

We have $(A_t^1, \ldots, A_t^D) \in \partial\Phi(Z_t)$, and

$$\begin{aligned}||(A_t^1, \ldots, A_t^D)|| \le& ((D-1)M + (1+D)\mu^{-1})||Z_t - Z_{t-1}||,\\ &\forall t > 0.\end{aligned} \quad (38)$$

Proof. Recalling (9), the optimal condition implies

$$\begin{aligned}\nabla_{X^d}\ell_t(X_{t-1}^d) + \mu^{-1}(X_t^d - X_{t-1}^d) + u_t^d = 0,\\ d = 1, \ldots, D,\end{aligned} \quad (39)$$

where $u_t^i \in \partial h_i(X_t^1)$. It is clear that

$$\begin{aligned}\nabla_{X^d}\ell_t(X_t^d) + u_t^d \in \partial_{X^d}\Phi(Z_t)\\ d = 1, \ldots, D,\end{aligned} \quad (40)$$

then we can conclude $(A_t^1, \ldots, A_t^D) \in \partial\Phi(Z_t)$.

Based on Assumption **(A4)** and assuming that the sequence $\{Z_t\}_{t\in\mathbb{N}}$ is bounded, for $d = 1, \ldots, D-1$ we have

$$\begin{aligned}||A_t^d|| \le& \mu^{-1}||X_{t-1}^d - X_t^d|| + ||\nabla_{X^d}\ell_t(Z_t) - \nabla_{X^d}\ell_t(Z_{t-1})||\\ \le& \mu^{-1}||X_{t-1}^d - X_t^d|| + M||Z_t - Z_{t-1}||\\ \le& (M + \mu^{-1})||X_{t-1}^d - X_t^1|| + M\sum_{d'\ne d}||X_{t-1}^{d'} - X_t^{d'}||\\ \le& (M + \mu^{-1})||Z_{t-1} - Z_t||,\end{aligned} \quad (41)$$

where we use the fact that $\nabla\ell$ is M-Lipschitz continuous on bounded subsets. For $d = D$, following the Lipschitz continuous gradient property of X^D and the fact that $\mu^{-1} \ge L_D$, we have

$$\begin{aligned}||A_t^D|| \le& \mu^{-1}||X_{t-1}^D - X_t^D|| + ||\nabla_{X^D}\ell(X_{t-1}^D) - \nabla_{X^D}\ell(X_t^D)||\\ \le& \mu^{-1}||X_{t-1}^D - X_t^D|| + \mu^{-1}||X_{t-1}^D - X_t^D||\\ \le& 2\mu^{-1}||X_{t-1}^D - X_t^D||.\end{aligned} \quad (42)$$

When $t > 0$, we can conclude

$$\begin{aligned}||(A_t^1, \ldots, A_t^D)|| \le& \sum_{d=1}^{D}||A_t^d||\\ \le& ((D-1)M + (D+1)\mu^{-1})||Z_t - Z_{t-1}||.\end{aligned} \quad (43)$$

□

By modifying the two lemmas above, we can conclude the properties of the limit point set. Let $\{Z_t\}_{t\in\mathbb{N}}$ be the sequence generated by Algorithm 2 from Z_0. The set of all limit points is denoted by

$$\begin{aligned}\text{limit}(Z_0) = \{\hat{Z} \in \mathbb{R}^{n_1\times m_1} \times \ldots \times \mathbb{R}^{n_D\times m_D} : \\ \exists \text{ an increasing sequence of integers } \{t_l\}_{l\in\mathbb{N}}, \\ Z^{t_l} \to \hat{Z} \text{ as } t_l \to \infty\}.\end{aligned} \tag{44}$$

Lemma 5. (Properties of limit(Z_0)) [25] *Suppose that assumptions **(A2)** and **(A4)** are satisfied. Let $\{Z_t\}_{t\in\mathbb{N}}$ be the sequence generated by Algorithm 2 with start point Z_0. The following assertions hold.*

(i) $\varnothing \neq \text{limit}(Z_0) \subset \text{crit}\ (\Phi)$*, where* crit (Φ) *is the set of critical points of* Φ.

(ii) We have

$$\lim_{t\to\infty} \text{dist}(Z_t, \text{limit}(Z_0)) = 0. \tag{45}$$

(iii) limit(Z_0) *is a non-empty, compact and connected set.*

(iv) The objective Φ *is finite and constant on* limit(Z_0).

This lemma follows the demonstration that both the upper and lower bounds go to the same limit point where the gradient is $\mathbf{0}$, which indicates (i) and (ii). Then by viewing $\text{limit}(Z_0)$ as an intersection of non-empty compact sets, we can validate (iii). Since the convergence of the sequence has been proved (in assertion (i)) already, we can obtain (iv). The explicit proof of this lemma could be found in [25], we omit the details here.

In the end, we can conclude the convergence properties in the multivariate scenario in the following theorem.

Theorem 2. (Main) *If assumptions **(A2)**, **(A3)** and **(A4)** hold, a step size is chosen such that $\mu < 1/L_{\max}$ where $L_{\max}$ is the maximum of $\{L_d\}_{d=1,\ldots,D}$, then the sequence $\{(X_t^1, \ldots, X_t^D)\}_{t\in\mathbb{N}}$ generated by any alternative proximal gradient method, such as Algorithm 2, will have finite length and converge to a critical point of (7). That is*

(i) The sequence $\{Z_t\}_{t\in\mathbb{N}}$ has finite length,

$$\sum_{t=1}^{\infty} ||Z_{t+1} - Z_t|| < \infty \tag{46}$$

(ii) The sequence $\{Z_t\}_{t\in\mathbb{N}}$ converges to a critical point Z^* of (7).

Proof. Based on Lemma 1 to Lemma 4, We can see that all conditions of Theorem 1 in [25] are satisfied. By extending the proof to the multivariate case, we can complete the proof here. □

The next thing we need to prove is that Φ has the KL property. Following [23] [25] [26], the proper and lower semi-continuous function σ will satisfy the KL property at any point of their domains, given σ is semi-algebraic. This is a sufficient condition of the KL property. The family of semi-algebraic functions could be summarized as follows.

Definition 2. (Semi-algebraic sets and functions)

(i) A subset S of $\mathbb{R}^n$ is a real semi-algebraic set if there exists a finite number of polynomial functions $g_{ij}, g'_{ij} : \mathbb{R}^n \to \mathbb{R}$ such that

$$S = \cup_{j=1}^{p} \cap_{i=1}^{q} \{u \in \mathbb{R}^n, g_{ij}(u) = 0 \text{ and } g'_{ij}(u) < 0\} \tag{47}$$

(ii) A function $r : \mathbb{R}^n \to (-\infty, +\infty]$ is called semi-algebraic if its graph

$$\{(u, \xi) \in \mathbb{R}^{n+1} : r(u) = \xi\} \tag{48}$$

is a semi-algebraic subset of $\mathbb{R}^{n+1}$.

It is worth mentioning that the definitions above could be extended to $\mathbb{R}^{n\times m}$. Following the definitions, we can prove that our objective Φ has the KL property.

Theorem 3. *An objective Φ defined in (4) with a penalty defined in (5) satisfies the KL property.*

Proof. As we can see, ℓ defined in (4) can be viewed as the sum of quadratic functions which are semi-algebraic. Hence ℓ is also semi-algebraic, and its graph in $\mathbb{R}^{n_1\times m_1} \times \ldots \times \mathbb{R}^{n_D\times m_D} \times \mathbb{R}$ is

$$\{(X^1, \ldots, X^D, \xi) \in \mathbb{R}^{n_1\times m_1} \times \ldots \times \mathbb{R}^{n_D\times m_D} \times \mathbb{R}_+ : \ell - \xi = 0\} \tag{49}$$

When h_d is defined as (5), we first investigate the auxiliary function $h' : \mathbb{R}^{n\times m_d} \times \mathbb{R}^{n\times k} \times \mathbb{R}^{m_d\times k} \to \mathbb{R}$, which satisfies $h'(X, U, V) = h(X)$, $U^\top U = I$ and $V^\top V = I$. Its graph in $\mathbb{R}^{n\times m_d} \times \mathbb{R}^{n\times k} \times \mathbb{R}^{m_d\times k} \times \mathbb{R}$ can be written as

$$\begin{aligned}\{(X, U, V, t) :\ \sigma_i \in \mathbb{R}_+, X = U\text{diag}(\{\sigma_i\})V^\top, \\ U^\top U - I = 0,\ V^\top V - I = 0, \text{ and } \sum_{i=1}^{k} w_i\sigma_i - \xi = 0\}\end{aligned} \tag{50}$$

We can see that the graph of h' in the subspace $\mathbb{R}^{n\times m_d}$ is exactly the graph of h. Base on Definition 2, (50) is a semi-algebraic set. Then following the Tarski-Seidenberg Theorem [27], the graph of h is also a semi-algebraic set, since its image can be obtained with the projection of a semi-algebraic set on the space of the first coordinate. It is obvious that the polynomial functions describing the graph of Φ is the sum of polynomial functions describing the graph of ℓ and (50) for all h_d. Thus the graph of Φ a semi-algebraic subset of $\mathbb{R}^{n_1\times m_1} \times \ldots \times \mathbb{R}^{n_D\times m_D} \times \mathbb{R}$. This completes the proof. □

It is clear that in (4), $\{\nabla_{X^d}\ell\}$ for all $d = 1, \ldots, D$ are Lipschitz continuous and $M = D$. Then following Theorem 2 and Theorem 3, we can get following conclusion.

Corollary 3. *Algorithm 2 will converge to a critical point in finite steps when solving problem (4).*

Corollary 2 shows that the GSVR-PG algorithm can achieve $O(1/T)$ sub-linear convergence rate for problem (4) under general conditions. For aligned matrix problem, if we choose $\{h_d\}$ as (5), it is easy to verify that, in each iteration, the extra computational complexity of introducing a shared matrix will be 1 to D times the cost of solving the problems in each domain separately, which is $O(Dnm\max\{\text{rank}(X_d)\})$, linear

regarding the number of domains in the worst case. Besides, following Corollary 3, there exists a positive integer t_l such that Algorithm 2 will converge faster than $\Omega(1/T)$ when $t > t_l$. In practice, we find that both Algorithm 1 and Algorithm 2 converge almost linearly, which indicates they are practical for large scale problems.

C. Accelerated Computation for Large Datasets

The most time consuming part of the proximal method above is an SVD computation in each iteration, which makes its scalability an issue in real-world applications. To accelerate the convergence, we use line-search to choose $\mu_{(t)}$ instead of a constant step size. Specifically, one can increase $\mu_{(t)}$ by $\mu_{(t)} = \eta\mu_{(t-1)},\ \eta > 1$ and make sure the inequality

$$\ell(X^d_{(t+1)}) < \ell(X^d_{(t)}) - \sigma||X^d_{(t+1)} - X^d_{(t)}||^2, \sigma \in (0,1) \quad (51)$$

is strictly satisfied unless $\mu^{(t+1)} < 1/L_{\max}$. In the meantime, a larger step size would lead to fewer positive components when solving shrinkage-thresholding problems, which implies lower rank of $X^d_{(t+1)}$ and fewer singular values to compute. This strategy guarantees that (34) is satisfied and the convergence is still promised.

Furthermore, as we observe from the convergent sequence, the rank may start and decrease from a large number which entails inefficient computation at the beginning. We use a decreasing sequence $\{\tau_0, \ldots, \tau_l\}$ with $\tau_l \leq 1$ to reduce the number of singular values above the threshold. In each iteration, the proximal map is computed as $P_h^{\tau_{(t)}\mu_{(t)}}(M_{(t)})$. It is clear that the convergence property is not affected as $\{\tau_i\}$ is a finite sequence. In practice, we set $\tau_0 = 10^2$ and $\tau_{(i+1)} = \max\{1, 0.7 \times \tau_i\}$. Besides, stochastic SVD [28] is also a practical approach to compute singular values for large datasets.

V. Experiments

To evaluate our method of aligned matrix completion (Aligned MC), we conduct experiments on both synthetic data and the task of multi-domain recommendation. We compare with the following baselines including both traditional matrix completion approaches and recommendation methods:

- **SVT** [29], a traditional matrix completion method which minimizes the nuclear norm.
- **SVP** [9], a matrix completion method based on singular value projection.
- **TNNR** [8], a matrix completion method which optimizes the truncated nuclear norm.
- **PMF** [30], probabilistic matrix factorization which is also a classical collaborative filtering method.
- **CMF** [7], a matrix factorization method that decomposes multiple matrices jointly, assuming common latent factors for all the domains.
- **GSMF** [16], a group-sparse matrix factorization method that incorporates group sparsity on the latent factors across multiple domains.

To investigate the behavior of the proposed method, we also evaluate the performance of Aligned MC where the GSVR term in (4) is replaced by the standard nuclear norm (Aligned MC-NN). In addition, as aforementioned, the proposed GSVR-PG algorithm is applicable for a family of singular value regularization functions including the truncated nuclear norm (TNNR), therefore in the experiments we compare the solutions of TNNR produced by our proposed algorithm (TNNR-PG) and the original algorithm in [8] (TNNR-Original) on the synthetic data to evaluate the effectiveness of the optimization algorithms.

TABLE I
Statistics of the multi-domain recommendation data

Domains	Book	Movie
#Users	13090	13090
#Items	17590	17922
Sparsity	99.66%	98.68%

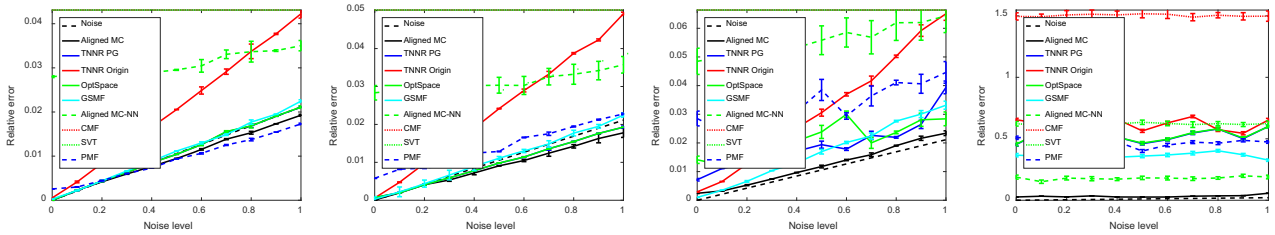
A. Synthetic Data

The synthetic data is constructed on two domains for experimental investigation. We randomly generate two 100×100 matrices with shared and distinct components as follows:

$$Z^d = M^d + D^d,\ \ Y^d_\Omega = Z^d_\Omega + \varepsilon,\ d = 1, 2. \quad (52)$$

Here $\{Z^d\}$ are the ground truth for all the domains, and $\{Y^d_\Omega\}$ are the noisy observed matrices. The shared components are generated by $M^d = AB^d$ where A is shared across all the domains, $A \in \mathbb{R}^{100\times 10}$ and $B^d \in \mathbb{R}^{10\times 100}$ consist of i.i.d. Gaussian entries with variance 25. The distinct parts are generated by $D^d = P^dQ^d$ where $P^d \in \mathbb{R}^{100\times 10}$ and $Q^d \in \mathbb{R}^{10\times 100}$ also consist of i.i.d. Gaussian entries but with variance 100. The observation indexes $\{\Omega_d\}$ are sampled uniformly at random. The variance of the shared components is set smaller than that of the distinct components to simulate real situations. The measure of relative error $\mathrm{RE} = \sum_{d=1}^2 ||X^{d^*} - Z^d||/(\sum_{d=1}^2 ||Z^d||)$ is used to evaluate the quality of the recovered matrices X^{d^*}. We set the parameters pen_1 and pen_2 in (5) proportional to the noise level, $C_1\sigma$ and $C_2\sigma$ respectively, where constants $C_1,\ C_2 \geq 0$ and σ is the standard variance of noise, $\gamma = 20$ and $k = 10$ for both shared and distinct parts.

We run all algorithms 10 times to obtain the means and standard deviations of RE under each observed ratio and noise level. The results are shown in Figure 1. We can first observe that CMF and SVT fail to recover the matrices in all settings. The performance of CMF is likely due to the fact that the distinct components are more significant than the shared part, contradicting with the assumption of CMF; while the number of observed entries does not satisfy the recovery condition of SVT, which explains its degeneration of performance. On the other hand, the benefits of a shared latent structure with domain-specific patterns are verified by smaller RE values of Aligned MC-NN compared with SVT, especially when the observation ratio drops to 40%. Meanwhile, the improvement of Aligned MC over Aligned MC-NN justifies



(a) 70% observed (b) 60% observed (c) 50% observed (d) 40% observed

Fig. 1. Relative error versus noise with different observation ratios

TABLE II
COMPARISON OF PERFORMANCE WITH DIFFERENT TRAINING RATIOS. RESULTS ARE PRESENTED IN THE FORM OF $\mathrm{RMSE}_{\mathrm{TEST}}(\mathrm{RMSE}_{\mathrm{TRAIN}})$.

Domains	Training	SVP	TNNR-PG	PMF	CMF	GSMF	Aligned MC
Book	80%	0.9606(0.4898)	0.8801 (0.6144)	0.7809 (0.5235)	0.8172 (0.6362)	0.7813 (0.5684)	**0.7389 (0.4008)**
	60%	1.0147(0.4658)	0.9066 (0.5663)	0.7967 (0.5353)	0.8517 (0.6523)	0.7962(0.6078)	**0.7479 (0.4550)**
	40%	1.1571(0.4175)	1.0239 (0.5563)	0.8397 (0.5083)	0.9345 (0.6227)	0.8030 (0.5643)	**0.7558 (0.4911)**
Movie	80%	0.7661(0.6011)	0.7336 (0.6524)	0.7342 (0.6014)	0.7325 (0.6228)	0.7315 (0.6177)	**0.7130 (0.6367)**
	60%	0.7870(0.5905)	0.7429 (0.6391)	0.7432 (0.5952)	0.7423 (0.6142)	0.7401(0.5978)	**0.7209 (0.6643)**
	40%	0.8387(0.5616)	0.7752 (0.6259)	0.7678 (0.5764)	0.7829 (0.5784)	0.7870 (0.4892)	**0.7342 (0.6885)**

the advantage of the GSVR regularization over the standard nuclear norm. All the other algorithms perform reasonably when the observation ratio is above 60%. Comparing the results of TNNR-PG and TNNR-Original, we confirm the stability of our proposed optimization algorithm. When the ratio decreases to 50%, the RE values of all the baselines grow faster with increasing noise than Aligned MC. When the observed ratio drops to 40%, all the comparing methods fail to recover the matrices correctly even if the observations are noiseless; whereas Aligned MC is capable of exploit the correlations among multiple domains to significantly alleviate the sparsity problem, which justifies our motivation.

B. Multi-Domain Recommendation

To measure the performance of aligned MC in the practical task of multi-domain recommendation, we use the data from a public website Douban[1], where users can rate movies, books and music, etc. We take two domains of ratings, *books* and *movies* in our experiment. We remove users and items with less than 10 ratings to provide enough ratings for split into training and test sets for evaluation. A dataset is then obtained containing 13090 users with 17590 ratings on books and 17922 ratings on movies. All ratings take values from 1 to 5. The details of the dataset are listed in Table I.

To evaluate the quality of recommendation, we use Root Mean Square Error, $\mathrm{RMSE}(X) = \sqrt{||X_\Omega - Y_\Omega||^2/N}$, to measure the discrepancy of predictions and the ground truth. We compare to both matrix completion algorithms and recommendation methods here as well. The parameters of our algorithm are set as follows: for the shared part, $\mathrm{pen}_1 = 65$, $\mathrm{pen}_2 = 300$, $\gamma = 5$ and $k = 20$; for the distinct part, $\mathrm{pen}_1 = 45$, $\mathrm{pen}_2 = 300$, $\gamma = 5$ and $k = 30$. We conduct the experiments with different training ratios (80%, 60% and 40%) for a comprehensive comparison. The training sets are sampled uniformly at random and the procedure is repeated 10 times. The results are summarized in Table II, where test RMSE values are shown with training RMSE values inside the brackets. Bold values indicate the best performance on the test data that is statistically significant with 95% confidence. The results of SVT, TNNR-Original and Aligned MC-NN are not reported here because they have to compute more than 600 singular values in the first dozens of iterations which are too expensive to produce the results in time. On the other hand, TNNR-PG adopts the proposed Algorithm 1 and avoids exhaustively computing the smaller singular values, which shows the capability of our algorithms for large scale problems.

From Table II, we can observe that all the recommendation methods achieve comparable performance in the *movie* domain, which contains relatively sufficient training data. Meanwhile in the *book* domain, CMF does not perform very well as the training set is extremely sparse and the connection between domains is weaker than it assumes. The performance of GSMF, which allows different factors for different domains, is comparable to PMF, and better than the other baselines. TNNR-PG performs comparably with the recommendation methods in the *movie* domain, while in the *book* domain the performances of the matrix completion approaches degenerate significantly. This is probably because SVP and TNNR are more sensitive to noise when sparsity is high. The last column records the results of our proposed method of Aligned MC which demonstrates significant superiority over the comparing algorithms. This justifies that Aligned MC can effectively exploit the consistency while modeling independency across

[1]http://www.douban.com

multiple domains with the benefits of improving the quality of recommendation.

VI. Conclusion

In this paper we consider the problem of matrix completion in multiple domains where multiple related matrices are reconstructed simultaneously. We propose the method of Aligned Matrix Completion, where we maintain consistency among all the domains while allowing independency of each separate domain. A general weighted singular value regularization is introduced for low-rank matrix completion, with theoretical convergence guarantee. Empirical results on synthetic data and the multi-domain recommendation task validate the capability of Aligned MC to effectively recover the low-rank structure of matrices and exploit the correlations among multiple domains to alleviate the sparsity problem.

VII. Acknowledgements

Research supported by grants from the National Natural Science Foundation of China (No. 61375060 and No. 61673364), the National Key Research and Development Program of China (No. 2016YFB1000904), the National Science Foundation for Distinguished Young Scholars of China (No. 61325010) and the Science and Technology Program for Public Wellbeing (No. 2013GS340302).

References

[1] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, 2009.

[2] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, 1992.

[3] H. Ji, C. Liu, Z. Shen, and Y. Xu, "Robust video denoising using low rank matrix completion," in *Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[4] A. Goldberg, B. Recht, J. Xu, R. Nowak, and X. Zhu, "Transduction with matrix completion: Three birds with one stone," in *Advances in Neural Information Processing Systems 23*, 2010.

[5] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

[6] M. White, X. Zhang, D. Schuurmans, and Y.-L. Yu, "Convex multi-view subspace learning," in *Advances in Neural Information Processing Systems 25*, 2012.

[7] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.

[8] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, "Fast and accurate matrix completion via truncated nuclear norm regularization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 9, 2013.

[9] P. Jain, R. Meka, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," in *Advances in Neural Information Processing Systems*, 2010.

[10] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014.

[11] X. Zhong, L. Xu, Y. Li, Z. Liu, and E. Chen, "A nonconvex relaxation approach for rank minimization problems," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.

[12] B. Li, Q. Yang, and X. Xue, "Can movies and books collaborate?: Cross-domain collaborative filtering for sparsity reduction," in *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, 2009.

[13] W. Pan, N. N. Liu, E. W. Xiang, and Q. Yang, "Transfer learning to predict missing ratings via heterogeneous user feedbacks," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 2011.

[14] H. Jing, A. Liang, S. Lin, and Y. Tsao, "A transfer probabilistic collective factorization model to handle sparse data in collaborative filtering," in *Proceedings of the 2014 IEEE International Conference on Data Mining*, 2014.

[15] Y. Zhang, B. Cao, and D.-Y. Yeung, "Multi-domain collaborative filtering." in *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 2010.

[16] T. Yuan, J. Cheng, X. Zhang, S. Qiu, and H. Lu, "Recommendation by mining multiple user behaviors with group sparsity," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2014.

[17] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, 2010.

[18] Y. Zhang and Z. Lu, "Penalty decomposition methods for rank minimization," in *Advances in Neural Information Processing Systems*, 2011.

[19] A. S. Lewis, "The mathematics of eigenvalue optimization," *Mathematical Programming*, vol. 97, no. 1-2, pp. 155–176, 2003.

[20] K. Kurdyka, "On gradients of functions definable in o-minimal structures," in *Annales de l'institut Fourier*, vol. 48, no. 3, 1998, pp. 769–783.

[21] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the kurdyka-lojasiewicz inequality," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 438–457, 2010.

[22] Y. Yu, X. Zheng, M. Marchetti-Bowick, and E. P. Xing, "Minimizing nonconvex non-separable functions," in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.

[23] H. Attouch and J. Bolte, "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features," *Mathematical Programming*, vol. 116, no. 1-2, pp. 5–16, 2009.

[24] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, 2008.

[25] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.

[26] J. Bolte, A. Daniilidis, and A. Lewis, "The lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 1205–1223, 2007.

[27] M. Coste, *An introduction to semialgebraic geometry*. Citeseer, 2000.

[28] O. Shamir, "A stochastic pca and svd algorithm with an exponential convergence rate," in *Proc. of the 32st Int. Conf. Machine Learning (ICML 2015)*, 2015, pp. 144–152.

[29] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[30] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," in *Advances in Neural Information Processing Systems*, 2008.