Partial Collective Matrix Factorization and its PAC Bound

Chao Lan, Xiaoli Li, Yujie Deng, Jun Huan
Department of Electrical Engineering and Computer Science
University of Kansas, 1450 Jayhawk Blvd.
Lawrence, KS 66045

Abstract

Collective matrix factorization (CMF) is a popular technique that factorizes multiple matrices jointly to boost the overall factorization quality. However, it has been argued the CMF assumptions are too strong: input matrices share a same rank and a same factor. A few heuristics were proposed to relax either assumption, but no theoretical justification was given.

In this paper, we promoted a prior solution to the theoretical level and formalized an assumption-free model called partial collective matrix factorization. It is based on the fact that any two matrices (of the same row) admit some joint factorization such that their factors are partially shared. We imported the computational learning theory to analyze this model, and proved its PAC bound for the matrix recovery task. Under mild conditions, we also identified the optimal choice of parameters for the proposed model. Finally, we implemented a simple algorithm motivated by the proposed model, and our simulation results demonstrated its superiority.

Introduction

Collective Matrix Factorization (CMF) (Singh and Gordon 2008) is becoming a standard technique for boosting the overall factorization quality of multiple related matrices. See studies in (Yu, Yu, and Tresp 2005; Lippert et al. 2008; Hofmann 2001; Singh and Gordon 2010; Bouchard, Yin, and Guo 2013; Zhou et al. 2014; Ding, Guo, and Zhou 2014; Yang, Jing, and Ng 2015) for example. These factors can be later used for recovering missing values in the matrix, or as the latent feature of an instance set for set classification or set clustering. In its general form, CMF considers a finite set of low rank matrices $\{X_i \in \mathbb{R}^{n\times p_i}\}_{i\in\mathcal{I}}$ of the same row dimension, and jointly factorizes them based on the form

$$X_i = DA_i \text{ for all } i \in \mathcal{I}, \tag{1}$$

with the assumption that these matrices share a same *factor* $D \in \mathbb{R}^{n\times k}$ for some $k \ll \min\{n, p\}$, and each matrix has its own *loading* $A_i \in \mathbb{R}^{k\times p_i}$. The factor typically has full column rank, either by assumption or by algorithmic constraint, e.g. (Yu, Yu, and Tresp 2005; Tan et al. 2014).

It is noted that CMF has essentially adopted two (nested) assumptions, that is, all matrices share a same rank and all matrices share a same factor. However, as argued in prior works, neither assumption is easy to satisfy in reality. In (Agarwal, Chen, and Long 2011; Zhang, Cao, and Yeung 2010), authors relaxed the shared-factor assumption to that each matrix has its own factor $D_i \in \mathbb{R}^{n\times k}$ and all D_i's are related by being drawn from a same distribution. We notice this relaxation has implicitly maintained the shared-rank assumption by requiring factors to have the same dimension. In (Klami et al. 2014), authors pointed out the limitation of the shared-rank assumption and relaxed it to that each matrix has its own factor $D_i \in \mathcal{R}^{n\times k_i}$ and all factors are related by having their column vectors chosen from a same vector pool. While these efforts are plausible from a practical point of view, they are all heuristic and the theoretical nature of their solutions remains unknown to us.

In this paper, we promote (Klami et al. 2014) to a theoretical level, and formalized an *assumption-free* factorization model called partial collective matrix factorization (pCMF). Our primary observation is that any set $\{X_i \in \mathbb{R}^{n\times p_i}\}_{i\in\mathcal{I}}$ admits the following form of joint factorization

$$X_i = \bar{D}P_i + \tilde{D}_i Q_i \text{ for all } i \in \mathcal{I}, \tag{2}$$

where factor $\bar{D} \in \mathbb{R}^{n\times c}$ (with proper choice of c) is shared by all matrices and factor $\tilde{D}_i \in \mathbb{R}^{n\times k_i}$ may differ for different matrices. Then, we carefully cast the problem of matrix recovery [1] by pCMF into a PAC framework (Kearns and Vazirani 1994), and concluded that given a set of matrices and a random sample of their observed entries, any estimated factors that are able to exactly recover the observed entry set can, with high probability, well recover all entries of the matrices. Under mild conditions, we further identified an optimal choice of model parameters c and k_i's that yield the smallest error bound, that is, $k_i = rank(X_i) - c$ for all $i \in \mathcal{I}$ and $c = \sum_{i=1}^{m} rank(X_i) - rank([X_1, ..., X_{|\mathcal{I}|}])$, where $[X_1, X_2]$ denotes the row concatenation of X_1, X_2. Finally, we implemented a simple pCMF algorithm, whose simulation results supported our theoretical study.

Technically, we imported tools from computational learning theory into the analysis of factorization-based matrix recovery problem, whose analysis is dominated by algebra tools, e.g. (Gunasekar et al. 2015). During our investigation, two interesting challenges arose. First, since we evaluated the quality of a factor by its optimal recovery ability similar to (Maurer and Pontil 2010; Rudi, Canas, and Rosasco

[1] A matrix recovery task aims to accurately recover all entries in a matrix based on a (observed) subset of its entries.

2013), one cannot follow the standard PAC argument and simply decompose the optimal recovery error on a sample of entries into the multiplication of optimal recovery errors on each entry, which is always zero. Second, PAC assumes sampling with replacement whereas in matrix recovery one typically considers sampling without replacement. Although authors in (Srebro, Alon, and Jaakkola 2004) have argued that bounds obtained under these two settings shall not differ significantly, no specific treatment for sampling without replacement was presented in that paper, and here we could not apply their argument due to the first technical challenge. As will be shown later, we tackled both challenges by directly bounding recovery error using the definition of probability under mild assumptions. Another interesting technical treatment in our analysis is the conversion of a coupled two-matrix factorization problem into three independent matrix factorization problem, which we carefully realized by the nature of pCMF. Finally, we established a connection between pCMF and the sum and intersection property of vector space (Warner 1965), so as to identify the optimal choice of model parameters.

The rest of this paper is organized as follows: in section two, we introduce the notations, concepts and assumptions adopted in this paper; our theoretical results are presented in section three, followed by an implementation of the pCMF model and its simulation in section four; conclusions and future works are discussed in section five.

Preliminaries

In this section we introduce the notations, concepts and assumptions for presenting the main result. Additional notions will be introduced in the proof.

Let us begin with the basic factorization model. Fix an $n \in \mathbb{N}$ and for any $p \in \mathbb{N}$, let $\mathcal{X}_p$ be a set of n-by-p matrices and $\mathcal{I}_p := \{(i,j); i = 1,..,n; j = 1,..,p\}$ be its index set. The entry of X at row i and column j will be mainly denoted by $X_{i,j}$, and occasionally by $X(i,j)$ for neat representation. For any $X \in \mathcal{X}_p$, let $rank(X)$ denote its rank, $span(X)$ denote its column space and $dom(X) := \mathbb{R}^{n\times p}$. Let $||X||_0$ be its L_0-norm and for any index set $\upsilon \subseteq \mathcal{I}_p$, define notation

$$||X||_0^{\upsilon} := \sum_{(i,j)\in\upsilon} \mathbf{1}\{X(i,j) \neq 0\}, \tag{3}$$

where $\mathbf{1}\{E\}$ is an indicator function taking 1 if event E is true and taking 0 if E is false.

For any factorization of X that takes the form DA, we shall call $D \in \mathbb{R}^{n\times k}$ the *factor*, $A \in \mathbb{R}^{k\times p}$ the *loading* and assume $k \ll \min\{n,p\}$. Define equivalent class

$$[D] := \{D' \in dom(D); span(D') = span(D)\}, \tag{4}$$

which groups all factors that span the same subspace. To simplify analysis, we will focus on a finite set of equivalent classes by assuming the columns of all factors are drawn from a vector space of dimension n over a finite field $\mathbb{F}_d$. This assumption can be satisfied if the matrix entries are defined over a finite set, as in the application of recommendation system. Our assumption can also be viewed as a realization of the finite hypothesis space assumption adopted in PAC theory. Now, define

$$\mathcal{D}_r := \{[D]; D \in \mathbb{R}^{n\times r}, rank(D) = r\} \tag{5}$$

as the collection of all possible equivalent classes. Then by (Prasad 2010) its cardinality $|\mathcal{D}_r|$ is the Gaussian binomial coefficient $\binom{n}{r}_d$. Note if D is full rank, then all $D' \in [D]$ are full rank. For conciseness, let us slightly abuse the notation so that $D \in \mathcal{D}_r$ means $D \in [D]$ for some $[D] \in \mathcal{D}_r$.

We will focus on two-matrix factorization for clarity, but our results can be easily generalized to multi-matrix setting.

Definition 1. *For any $X_1 \in \mathcal{X}_{p_1}$ and $X_2 \in \mathcal{X}_{p_2}$, we say they admit a (c,k_1,k_2)-factorization if there exists factors $\bar{D} \in \mathcal{D}_c$, $\tilde{D}_1 \in \mathcal{D}_{k_1}$ and $\tilde{D}_2 \in \mathcal{D}_{k_2}$ such that*

$$X_1 = \bar{D}P_1 + \tilde{D}_1 Q_1 \quad \text{and} \quad X_2 = \bar{D}P_2 + \tilde{D}_2 Q_2 \tag{6}$$

for some associated loadings $P_1 \in \mathbb{R}^{c\times p_1}$, $P_2 \in \mathbb{R}^{c\times p_2}$, $Q_1 \in \mathbb{R}^{k_1\times p_1}$ and $Q_2 \in \mathbb{R}^{k_2\times p_2}$.

The (c,k_1,k_2)-factorization will serve as the basic factorization model in our analysis. By rewriting its definition as

$$X_1 = \left[\bar{D}, \tilde{D}_1\right]\begin{bmatrix}P_1\\Q_1\end{bmatrix} \text{ and } X_2 = \left[\bar{D}, \tilde{D}_2\right]\begin{bmatrix}P_2\\Q_2\end{bmatrix}, \tag{7}$$

it looks as if the factorization assumes two matrices partially share their factors. In particular, by fixing $k_1 = k_2 = 0$, we have $X_1 = \bar{D}P_1$ and $X_2 = \bar{D}P_2$, which is the factorization model adopted by CMF. The following fact shows the generality of our model and the limitation of CMF's.

Fact 2. *Given any two matrices $X_1 \in \mathcal{X}_{p_1}$ of rank r_1 and $X_2 \in \mathcal{X}_{p_2}$ of rank r_2, there always exists some $c,k_1,k_2 \geq 0$ satisfying $c+k_1 = r_1$ and $c+k_2 = r_2$, such that X_1 and X_2 admit a (c,k_1,k_2)-factorization. However, this conclusion may not be true if we fix $k_1 = k_2 = 0$.*

The last statement holds trivially when $r_1 \neq r_2$. Even when $r_1 = r_2$, the conclusion is still invalid if two matrix factors $\bar{D}$'s are drawn from different equivalent classes in $\mathcal{D}_c$.

In rest of the discussion, we will always assume $t \in \{1,2\}$ when it is used to index the two matrices X_1, X_2 or their parameters such as index sets, factors and loadings.

Notions related to learning

So far we have used notations $\bar{D}, \tilde{D}$ (or, A, P, Q) to represent factors (or, loadings) that can exactly recover matrices. Since these 'ideal' terms do not appear often in later analysis, we will overload their notations to represent the *estimate* of them in rest of the paper. The notations of 'ideal' terms will be additionally introduced whenever used.

Let $\upsilon_t \subseteq \mathcal{I}_{p_t}$ be a sub-index set, whose elements are assumed to be sampled uniformly from $\mathcal{I}_{p_t}$, but not necessarily independently. This is an assumption adopted in many analysis such as (Srebro, Alon, and Jaakkola 2004; Candès and Recht 2009). Notice that υ_t naturally induces a subset S_t of X_t where $S_t(i,j) = X_t(i,j)$ for all $(i,j) \in \upsilon_t$. For this reason, we will simply call υ_t the *sample* of X_t, and S_t the corresponding *sample entries*.

Definition 3. *For any matrices $X_1 \in \mathcal{X}_{p_1}$ and $X_2 \in \mathcal{X}_{p_2}$ and their respective samples υ_1 and υ_2, a (c, k_1, k_2)-partial CMF learner is an algorithm that, based on υ_t's, estimates a (c, k_1, k_2)-factorization admitted by X_1, X_2, and returns a bag of estimated factors $\hat{\Theta} = (\bar{D}, \tilde{D}_1, \tilde{D}_2)$. We say $\hat{\Theta}$ is consistent on υ_t's if there are P_t, Q_t's such that,*

$$||X_t - (\bar{D}P_t + \tilde{D}_t Q_t)||_0^{\upsilon_t} = 0 \quad \textit{for } t = 1, 2. \tag{8}$$

The quality of $\hat{\Theta}$ will be evaluated by its *ability* for matrix recovery. To this end, define the *matrix recovery error* as

$$er(\hat{\Theta}) := \inf_{P_t, Q_t} \sum_t ||X_t - (\bar{D}P_t + \tilde{D}_t Q_t)||_0 / |\mathcal{I}_+|, \tag{9}$$

where $|\mathcal{I}_+| = |\mathcal{I}_{p_1}| + |\mathcal{I}_{p_2}|$. This recovery error is essentially counting the number of entries incorrectly recovered by $\hat{\Theta}$, and we choose its form to facilitate the import of PAC theory. It is noted if X_t is binary in $\{0, 1\}$, then $||X_t||_0 = ||X_t||_F$ and our error becomes a variant of the reconstructive error in (Maurer and Pontil 2010; Rudi, Canas, and Rosasco 2013).

Our analysis also involves singular matrix factorization, whose notions largely parallel the ones we defined for CMF. An $X \in \mathcal{X}_p$ is said to admit a *k-factorization* if $X = \ddot{D}A$ for some $\ddot{D} \in \mathcal{D}_k$ and $A \in \mathbb{R}^{k \times p}$. Let $\upsilon \subseteq \mathcal{I}_p$ be a sample of X. A *k-SMF learner* estimates a k-factorization admitted by X based on υ, and returns an estimated factor D. The factor is said to be consistent on υ if $||X - DA||_0^{\upsilon} = 0$ for some A. The quality of D is evaluated by the matrix recovery error

$$er(D) := \inf_A ||X - DA||_0 / |\mathcal{I}_p|. \tag{10}$$

Main Result

Our main result is stated as follows.

Theorem 4. *For any matrices $X_1 \in \mathcal{X}_{p_1}$ and $X_2 \in \mathcal{X}_{p_2}$ and their respective random samples υ_1 and υ_2, let $\hat{\Theta}$ be a bag returned by the (c, k_1, k_2)-pCMF learner. Then with probability [2] at least $1 - \delta$, for all $\hat{\Theta}$ consistent on υ_t's, we have $er(\hat{\Theta}) \leq U(c, k_1, k_2)$, where*

$$\begin{aligned} U(c, k_1, k_2) = &\frac{1}{|\upsilon_1|} \log \binom{n}{k_1}_d + \frac{1}{|\upsilon_2|} \log \binom{n}{k_2}_d \\ &+ \frac{1}{|\upsilon_+|} \log \binom{n}{c}_d \\ &+ \left(\frac{1}{|\upsilon_+|} + \frac{1}{|\upsilon_1|} + \frac{1}{|\upsilon_2|} \right) \log \frac{3}{\delta}, \end{aligned} \tag{11}$$

and $|\upsilon_+| = |\upsilon_1| + |\upsilon_2|$. Further, let the optimal parameter be $(c^*, k_1^*, k_2^*) := \arg\min_{(c,k_1,k_2)} U(c, k_1, k_2)$. If $c, k_1, k_2 \leq n/2$, then $c^* = rank(X_1) + rank(X_2) - rank([X_1, X_2])$ and $k_t^* = rank(X_t) - c^*$ for $t = 1, 2$.

Let us interpret the results in Theorem 1 here. First, the bound has clearly inherited properties from the standard PAC bound: we see U decreases as the 'training sample size' $|\upsilon_1|$ and $|\upsilon_2|$ increase, or as 'confidence parameter' δ decreases. A seemingly difference is that U has weaker asymptotic guarantee with respect to $|\upsilon_1|$ and $|\upsilon_2|$, since both numbers are bounded from above, partly due to the scheme of sampling without replacement. This is in contrast to standard PAC theory where the training sample can grow arbitrarily large, partly due to its sampling with replacement scheme. However, we also notice this difference does not degrade the value of U, since our demand on the error bound has also become looser, i.e. for precise recovery it suffices to bound $er(\hat{\Theta})$ by $1/|\mathcal{I}_{p_t}|$ instead of an arbitrarily small ϵ. Another observation is U decreases as the first three logarithm terms increase. These terms arouse by a union bound over all possible equivalent classes in $\mathcal{D}_c$, $\mathcal{D}_{k_1}$ and $\mathcal{D}_{k_2}$ respectively, and correspond to the *size of hypothesis space* in PAC theory.

In addition to traditional implications, we also gained new insights from Theorem 1, mainly through the identification of optimal model parameters. Under mild conditions, we see a trade-off between c and k_t's, and that a larger c and smaller k_t's yields a smaller bound. This means the more partial sharing between matrix factors, the better performance guarantee we can obtain for the pCMF model. We also see how shared factor helps, i.e. the error contributed by shared factor is more down-weighted by larger sample size $|\upsilon_+|$.

Justification of Theorem 1

In this section we prove Theorem 1. Our investigation began with casting a single matrix recovery problem into the PAC framework and proved its PAC bound, where we tackled the challenges that υ is not sampled with replacement and the sample recovery error cannot be simply decomposed into the recovery error of each entry as PAC does. We then carefully re-cast the original two-matrix recovery problem into three decoupled single matrix recovery problems by a considerable amount of technical treatments, and applied earlier analysis three times to yield the error bound for the pCMF learner. Finally, we identified the optimal choice of model parameters under mild conditions.

Bounding the error of a SMF learner

Recall the notations defined for single matrix factorization in earlier section. For technical reasons, we will need to consider the choice of loading and thus temporarily expand the definition of $er(D)$ to

$$er_A(D) := ||X - DA||_0 / |\mathcal{I}_p|. \tag{12}$$

Clearly, $er(D) = \inf_A er_A(D)$. To facilitate discussion, let us additionally define the *sample recovery error rate* as

$$\hat{er}(D) := \inf_A \hat{er}_A(D), \tag{13}$$

where $\hat{er}_A(D) := ||X - DA||_0^{\upsilon} / |\upsilon|$.

Note D is consistent on sample υ if $\hat{er}(D) = 0$. Then our PAC bound for the k-SMF learner is stated as follows.

Lemma 5. *For any matrix $X \in \mathcal{X}_p$ and its sample υ, let D be a factor returned by the k-SMF learner. For any $\epsilon, \delta > 0$ and all D consistent on υ, with probability at least $1 - \delta$, we*

[2] The probability is taken over the random choice of υ_1, υ_2.

have $er(D) \leq \epsilon$ *if* $|\upsilon| \geq \frac{1}{\epsilon}\left(\log \binom{n}{k}_d + \log \frac{1}{\delta}\right)$. *Further,*

$$er(D) \leq \frac{1}{|\upsilon|}\left(\log \binom{n}{k}_d + \log \frac{1}{\delta}\right). \tag{14}$$

with probability [3] *at least* $1-\delta$.

Proof. The backbone of our proof follows standard PAC arguments. We also take care of the additional challenges that sample recovery error can no longer be simply decomposed and sample is no longer drawn with replacement.

Let us first fix a $D \in \mathcal{D}_k$ and the size of υ. Consider such a sample $\upsilon \subseteq \mathcal{I}_p$ that, even if D is able to correctly recover all entries in υ (i.e. $\hat{er}_{A'}(D) = 0$ for some A'), it still fails on at least a ϵ-fraction of the entire entry set $\mathcal{I}_p$ (i.e. $er(D) > \epsilon$). The chance of getting such misleading sample is

$$\begin{aligned}
&\Pr\{\upsilon \in \mathcal{I}_p; \hat{er}_{A'}(D) = 0 \wedge er(D) > \epsilon\} \\
&\leq \Pr\{\upsilon \in \mathcal{I}_p; \hat{er}_{A'}(D) = 0 \wedge er_{A'}(D) > \epsilon\} \\
&\leq \Pr\{\upsilon \in \mathcal{I}_p; \hat{er}_{A'}(D) = 0 \mid er_{A'}(D) > \epsilon\} \\
&\leq \binom{(1-\epsilon)|\mathcal{I}_p|}{|\upsilon|} \Big/ \binom{|\mathcal{I}_p|}{|\upsilon|},
\end{aligned} \tag{15}$$

where the first inequality is by the definition of $er(D)$, and the last inequality is by the uniform sampling assumption and the definition of probability. To be specific, the last probability counts the possible choice of υ whose entries are all correctly recovered, given that at most $(1-\epsilon)|\mathcal{I}_p|$ entries can be correctly recovered. Clearly, this count reaches its maximum when D correctly recovers exactly $(1-\epsilon)|\mathcal{I}_p|$ entries, leaving the maximum number of choices for υ, that is, $(1-\epsilon)|\mathcal{I}_p|$ chooses $|\upsilon|$. The denominator is simply a normalizer so that the result remains a valid probability mass.

It is noted we could not simply decompose $\hat{er}(D) = 0$ into the sample recovery errors of each entry in υ as in PAC analysis, because the latter error will always be zero by any factor, leaving no uncertainty to bound. It is also noted independent sampling assumption is not needed to derive (15).

Formula (15) applies to only one factor, and our next step is to go through all possible factors in $\mathcal{D}_k$ to obtain a uniform bound. Note this is equivalent to go through one factor per equivalent class $[D]$ and then through all possible classes, because all $D \in [D]$ share exactly the same sample/matrix recovery error on any υ, achieved with possibly different loadings. For convenience, define the bad event $\mathfrak{B}(D_i) := \{\hat{er}_{A'_i}(D_i) = 0 \wedge er(D_i) > \epsilon\}$. Then

$$\begin{aligned}
&\mathbb{P}\{\upsilon; \mathfrak{B}(D_1) \vee \mathfrak{B}(D_2) \vee ...\} \\
&= \mathbb{P}\{\upsilon; \mathfrak{B}(D_{i_1} \in [D]_1) \vee \mathfrak{B}(D_{i_2} \in [D]_2) \vee ...\} \\
&\leq \sum_{j=1}^{|\mathcal{D}_k|} \mathbb{P}\{\upsilon; \mathfrak{B}(D_{i_j} \in [D]_j)\} \\
&\leq \binom{n}{k}_d \binom{(1-\epsilon)|\mathcal{I}_p|}{|\upsilon|} \Big/ \binom{|\mathcal{I}_p|}{|\upsilon|}.
\end{aligned} \tag{16}$$

It remains to bound the right-most term in (16) by δ and solve for ϵ. Since direct solution is difficult to find, we rely on the relaxation that if $p \geq p' \geq \ell$, then $\binom{p'}{\ell}/\binom{p}{\ell} \leq (\frac{p'}{p})^{\ell}$. Then $\binom{(1-\epsilon)|\mathcal{I}_p|}{|\upsilon|}/\binom{|\mathcal{I}_p|}{|\upsilon|} \leq (1-\epsilon)^{|\upsilon|} \leq e^{-\epsilon|\upsilon|}$, and it suffices to solve $\binom{n}{k}_d e^{-\epsilon|\upsilon|} \leq \delta$ for ϵ, which proves the lemma. □

Let us briefly interpret the result in Lemma 1. Similar to standard PAC bound, it says a larger sample (i.e. bigger $|\upsilon|$) yields a lower error bound. A new feature in our result is the introduction of parameter k through $\binom{n}{k}_d$. Since this coefficient is small when k is either very small or very large, it seems both high rank and low rank factors help to improve the recovery performance. However, we point out that the option of high rank factor makes sense here only because we have abstracted away the influence of loading estimation by the definition of recovery error in (10). In practice, if the matrix rank is very low, choosing a high rank factor will introduce a great amount of redundant parameters that may cause over-fitting. Thus a more practical option is to choose low rank factors, which is widely seen in research.

It is also noted that, by the definition of k-SMF learner we have $k \geq rank(X)$, otherwise no k-factorization can be admitted by X. Although our bounding technique also applies to the case $k < rank(X)$, the resulted bound may not be very meaningful in a PAC sense, i.e. the bound would hold not mainly because it is rare that consistent bag has high recovery error, as we wish to conclude, but largely because it is rare that consistent bags can be returned.

Bounding the error of a pCMF learner

In this section we prove the error bound in Theorem 1. Let us begin with a primary observation that, under the pCMF model, a two coupled matrix factorization problem can be re-cast into three independent latent matrix factorization problems. Specifically, for any $X_1 \in \mathcal{X}_{p_1}$ and $X_2 \in \mathcal{X}_{p_2}$, by (6) and Fact 1 there always exist decompositions

$$X_1 = \bar{X}_1 + \tilde{X}_1 \quad \text{and} \quad X_2 = \bar{X}_2 + \tilde{X}_2, \tag{17}$$

so that estimating their admitted (c, k_1, k_2)-factorization appears as if we are estimating the c-rank factorization of $\bar{X}_1, \bar{X}_2$ using $\bar{D}$, and estimating the k_t-rank factorization of $\tilde{X}_t$ using $\tilde{D}_t$. In this spirit, we have

$$\begin{aligned}
&||X_t - (\bar{D}P_t + \tilde{D}_t Q_t)||_0 \\
&= ||(\bar{X}_t - \bar{D}P_t) + (\tilde{X}_t - \tilde{D}_t Q_t)||_0 \\
&\leq ||\bar{X}_t - \bar{D}P_t||_0 + ||\tilde{X}_t - \tilde{D}_t Q_t||_0,
\end{aligned} \tag{18}$$

and thus overall

$$\begin{aligned}
&\sum_t ||X_t - \bar{D}P_t - \tilde{D}_t Q_t||_0 \\
&\leq ||\bar{X}_1 - \bar{D}P_1||_0 + ||\tilde{X}_1 - \tilde{D}_1 Q_1||_0 \\
&\quad + ||\bar{X}_2 - \bar{D}P_2||_0 + ||\tilde{X}_2 - \tilde{D}_2 Q_2||_0 \\
&= ||\tilde{X}_1 - \tilde{D}_1 Q_1||_0 + ||\tilde{X}_2 - \tilde{D}_2 Q_2||_0 \\
&\quad + ||\left[\bar{X}_1, \bar{X}_2\right] - \bar{D}\left[P_1, P_2\right]||_0.
\end{aligned} \tag{19}$$

Formula (19) seems to be suggesting the task of bounding the recovery error over two input matrices may be relaxed to

[3] The probability is taken over the random choice of υ.

the task of bounding the recovery error over three independent hidden matrices, namely, $\tilde{X}_1$, $\tilde{X}_2$ and $\left[\bar{X}_1, \bar{X}_2\right]$. Now, although the goal is clear, to realize such relaxation under the PAC framework involves a considerable amount of technical treatments, which are presented in rest of this section.

Let us firs introduce extra notions to facilitate discussion. Let p_υ denote a decomposition (depending on υ) of X_1, X_2 by the form (17). For any p_υ, define related recovery errors

$$\begin{aligned}
er_p^\upsilon(\tilde{D}_t) &:= \inf_{Q_t} ||\tilde{X}_t - \tilde{D}_t Q_t||_0 \,/\, |\mathcal{I}_t|,\\
er_p^\upsilon(\bar{D}) &:= \inf_{P_t' s} || \left[\bar{X}_1, \bar{X}_2\right] - \bar{D}\left[P_1, P_2\right] ||_0 \,/\, |\mathcal{I}_+|,\\
\hat{er}_p^\upsilon(\tilde{D}_t) &:= \inf_{A_t} ||\tilde{X}_t - \bar{D}A_t||_0^{\upsilon_t} \,/\, |\upsilon_t|,\\
\hat{er}_p^\upsilon(\bar{D}) &:= \inf_{A_t' s} \sum_t ||\bar{X}_t - \bar{D}A_t||_0^{\upsilon_t} \,/\, |\upsilon_+|,\\
\hat{er}(\hat{\Theta}) &:= \inf_{P_t, Q_t} \sum_t ||X_t - \bar{D}P_t - \tilde{D}_t Q_t||_0^{\upsilon_t} \,/\, |\upsilon_+|,
\end{aligned} \tag{20}$$

where $|\upsilon_+| = |\upsilon_1| + |\upsilon_2|$.

Our first result is a connection between the input matrices and the hidden matrices in terms of matrix recovery errors.

Lemma 6. *For any p_υ, we have*

$$er(\hat{\Theta}) \le er_p^\upsilon(\bar{D}) + er_p^\upsilon(\tilde{D}_1) + er_p^\upsilon(\tilde{D}_2). \tag{21}$$

Proof. By taking infimum over P_t, Q_t's, we have

$$\begin{aligned}
er(\hat{\Theta}) &= \inf \sum_t ||X_t - \bar{D}P_t - \tilde{D}_t Q_t||_0 \,/\, |\mathcal{I}_+|\\
&\le \inf(||\tilde{X}_1 - \tilde{D}_1 Q_1||_0 + ||\tilde{X}_2 - \tilde{D}_2 Q_2||_0\\
&\qquad + || \left[\bar{X}_1, \bar{X}_2\right] - \bar{D}\left[P_1, P_2\right] ||_0) \,/\, |\mathcal{I}_+|\\
&= \inf_{P_t' s} || \left[\bar{X}_1, \bar{X}_2\right] - \bar{D}\left[P_1, P_2\right] ||_0 \,/\, |\mathcal{I}_+|\\
&\quad + \inf_{Q_1} ||\tilde{X}_1 - \tilde{D}_1 Q_1||_0 \,/\, |\mathcal{I}_+|\\
&\quad + \inf_{Q_2} ||\tilde{X}_2 - \tilde{D}_2 Q_2||_0 \,/\, |\mathcal{I}_+|\\
&\le er_p^\upsilon(\bar{D}) + er_p^\upsilon(\tilde{D}_1) + er_p^\upsilon(\tilde{D}_2).
\end{aligned} \tag{22}$$

The first inequality is based on (19) and the last inequality is by relaxing $|\mathcal{I}_+| \ge |\mathcal{I}_t|$. Such relaxation is for simplifying discussion, and our final bound would remain the same without it, i.e., if instead we conclude $er(\hat{\Theta}) \le er_p^\upsilon(\bar{D}) + er_p^\upsilon(\tilde{D}_1)|\mathcal{I}_1|/|\mathcal{I}_+| + er_p^\upsilon(\tilde{D}_2)|\mathcal{I}_2|/|\mathcal{I}_+|$. To verify the second equality, we rely on the following result

Lemma 7. *(Wade 2003) Let A, B be nonempty sets and define set $A + B := \{a + b; a \in A \wedge b \in B\}$. Then* $\inf(A + B) = \inf A + \inf B$.

Now, the equality holds because the three infimum terms are independently taken over $\{P_1, P_2\}$, $\{Q_1\}$ and $\{Q_2\}$ respectively, and thus we can split them by above lemma. □

Based on Lemma 6, if $\epsilon = \epsilon_0 + \epsilon_1 + \epsilon_2$, then by the pigeonhole principle $er(\hat{\Theta}) > \epsilon$ implies at least one of the following: $er_p^\upsilon(\bar{D}) > \epsilon_0$ or $er_p^\upsilon(\tilde{D}_1) > \epsilon_1$ or $er_p^\upsilon(\tilde{D}_2) > \epsilon_2$. If we manage to bound the probability for each inequality over all consistent $\hat{\Theta}$ by $\delta/3$, then by a union bound we can bound $er(\hat{\Theta}) > \epsilon$ over all consistent $\hat{\Theta}$ by probability δ. Take $\tilde{D}_t$ for instance, for any p_υ define event

$$\tilde{\mathfrak{B}}_{p_\upsilon}^t(\hat{\Theta}) := \{\hat{er}(\hat{\Theta}) = 0 \wedge er_p^\upsilon(\tilde{D}_1) > \epsilon_1\}. \tag{23}$$

We wish to conclude that

$$\mathbb{P}\{\tilde{\mathbb{B}}_{p_\upsilon}^t(\hat{\Theta}_1) \vee \tilde{\mathbb{B}}_{p_\upsilon'}^t(\hat{\Theta}_2) \vee ...\} \le \delta/3, \tag{24}$$

where p_υ and p_υ' may be different. Our goal is to solve this inequality by applying the earlier results on single matrix recovery, so as to introduce parameters c, k_1 and k_2.

Notice any p_υ on matrix X_t naturally induces a decomposition on its sample υ_t so that $\bar{\upsilon}_t$ is the sample of $\bar{X}_t$ and $\tilde{\upsilon}_t$ is the sample of $\tilde{X}_t$. In essence, these samples have the same index set, but different sample entries.

Example 8. *Consider a p_υ on $X_t = [3, 5; 1, 7]$ that*

$$\begin{bmatrix} 3 & 5 \\ 1 & 7 \end{bmatrix} = \begin{bmatrix} 1 & 3.5 \\ 0.3 & 4 \end{bmatrix} + \begin{bmatrix} 2 & 1.5 \\ 0.7 & 3 \end{bmatrix} = \bar{X}_t + \tilde{X}_t. \tag{25}$$

Suppose the sample of X_t is $\upsilon_t = \{(1,1),(1,2)\}$, whose entries are $\{3, 5\}$. Then p_υ naturally induces a sample $\bar{\upsilon}_t = \{(1,1),(1,2)\}$ of $\bar{X}_t$ whose entries are $\{1,\ 3.5\}$ and a sample $\tilde{\upsilon}_t = \{(1,1),(1,2)\}$ of $\tilde{X}_t$ whose entries are $\{2,\ 1.5\}$.

Now we remark the following fact.

Fact 9. *For any $\hat{\Theta}$ consistent on υ_t's, there exists a p_υ such that $\bar{D}$ is consistent on $\bar{\upsilon}_1, \bar{\upsilon}_2$ and $\tilde{D}_t$ is consistent on $\tilde{\upsilon}_t$.*

The fact holds by constructing p_υ: for $\bar{X}_t, \tilde{X}_t$'s, their sample entries can be obtained by using $\hat{\Theta}$ to reconstruct $\bar{\upsilon}_t, \tilde{\upsilon}_t$'s where $\bar{D}, \tilde{D}_t$'s are consistent on accordingly; the rest entries can be obtained by arbitrary decomposition on X_t's.

A nice utility of Fact 2 is for any samples we can always choose that particular p_υ so that $\hat{er}(\hat{\Theta}) = 0$ implies $\hat{er}_p^\upsilon(\bar{D}) = 0$ and $\hat{er}_p^\upsilon(\tilde{D}_t) = 0$ for $t = 1, 2$. Then

$$\begin{aligned}
\mathbb{P}\{\tilde{\mathfrak{B}}_{p_\upsilon}^t(\hat{\Theta})\} &\le \mathbb{P}\{\hat{er}_p^\upsilon(\tilde{D}_t) = 0 \wedge er_p^\upsilon(\tilde{D}_t) > \epsilon_t\}\\
&\le \mathbb{P}\{\hat{er}_p^*(\tilde{D}_t) = 0 \wedge er_p^*(\tilde{D}_t) > \epsilon_t\}
\end{aligned} \tag{26}$$

where $er_p^*(\tilde{D}_t)$ is the recovery error based on a universal decomposition p_* that does not depend on samples. The second inequality holds because our constraint does not really consider the entry values (and thus which decomposed matrix/sample is). In essence, the probability does no more than counting the possible choices of υ_t whose indexing entries are all correctly recovered while overall there is at least an ϵ-fraction of entries incorrectly recovered, which is equivalent to our previous single matrix recovery problem. Similar argument has been applied in (Srebro, Alon, and Jaakkola 2004) by considering the 0-1 recovery loss function.

Define event $\tilde{\mathfrak{B}}_*^t(\tilde{D}_t) := \{\hat{er}_p^*(\tilde{D}_t) = 0 \wedge er_p^*(\tilde{D}_t) > \epsilon_t\}$

and let $\hat{\Theta}_i := \{\bar{D}_i, \tilde{D}_{1,i}, \tilde{D}_{2,i}\}$. Then we have

$$\begin{aligned}
&\mathbb{P}\{\tilde{\mathfrak{B}}^t_{p_\upsilon}(\hat{\Theta}_1) \vee \tilde{\mathfrak{B}}^t_{p'_\upsilon}(\hat{\Theta}_2) \vee ...\} \\
&\leq \mathbb{P}\{\tilde{\mathfrak{B}}^t_*(\tilde{D}_{t,1}) \vee \tilde{\mathfrak{B}}^t_*(\tilde{D}_{t,2}) \vee ...\} \\
&\leq \binom{n}{k_t}_d \mathbb{P}\{\tilde{\mathfrak{B}}^t_*(\tilde{D}_{t,1})\} \qquad (27) \\
&\leq \binom{n}{k_t}_d \binom{(1-\epsilon_t)|\mathcal{I}_p|}{|\upsilon_t|} \Big/ \binom{|\mathcal{I}_p|}{|\upsilon_t|}.
\end{aligned}$$

The first inequality follows our argument in (26), and second inequality follows a union bound and the fact that although we started with all possible choice of $\hat{\Theta}$, many of them have overlapping $\tilde{D}_t$ that can take at most $\binom{n}{k_t}_d$ distinct values.

Setting the right-most term in (27) less or equal to $\delta/3$ and solving for ϵ_t, we have for all consistent $\hat{\Theta}$ and $t = 1, 2$, with probability at least $1 - \delta/3$

$$er^\upsilon_p(\tilde{D}_t) \leq \frac{1}{|\upsilon_t|}\left(\log \binom{n}{k_t}_d + \log\frac{3}{\delta}\right). \qquad (28)$$

Following the same argument, we also conclude

$$er^\upsilon_p(\bar{D}) \leq \frac{1}{|\upsilon_+|}\left(\log \binom{n}{c}_d + \log\frac{3}{\delta}\right) \qquad (29)$$

for all consistent $\hat{\Theta}$ with probability at least $1 - \delta/3$.

Putting all together, recalling Lemma 6 and by a union bound proves the error bound in Theorem 1.

Identifying the optimal parameters

In this section, we show how to identify the optimal choice of parameters c and k_t's under mild conditions.

On one side, if $c, k'_t s \leq n/2$, then $U(c, k_1, k_2)$ is monotonically increasing with the parameters and thus one wishes to choose small c and k_t's. On the other side, recall that $c+k_t \geq rank(X_t)$ by the definition of pCMF learner, which guarantees the return of a consistent bag. As a result, the optimal choice is attained at $c + k_t = rank(X_t)$ for $t = 1, 2$, which reveals a trade-off between c and k_t's.

Note that the contributions of c and k_t to $U(c, k_1, k_2)$ are weighted differently, i.e. that of c is more down-weighted by $|\upsilon_+|$ and that of k_t is less down-weighted by $|\upsilon_t|$. Therefore, a small error bound is achieved at large c and small k_t.

It remains to identify the largest c (and the smallest k_t) that satisfies all above constraints. To this end, we notice a connection between our problem and the following result.

Lemma 10. *(Warner 1965) Let V_1 and V_2 be two subspaces of $\mathbb{R}^n$, and define $V_1 + V_2 := \{v + u; v \in V_1, u \in V_2\}$ and $V_1 \cap V_2 := \{v; v \in V_1 \wedge v \in V_2\}$. Then*

$$dim(V_1) + dim(V_2) = dim(V_1 + V_2) + dim(V_1 \cap V_2),$$

where $dim(V)$ denotes the dimension of subspace V.

According to Lemma 10, if we manage to map $dim(V_t)$ to $rank(X_t)$, map $dim(V_1 \cap V_2)$ to the largest c and map $dim(V_1 + V_2)$ to $rank([X_1, X_2])$, then we will prove the optimal choice of parameters in Theorem 1. The next corollary is based on a realization of these mappings.

Corollary 11. *For any $X_1 \in \mathcal{X}_{p_1}$ of rank r_1 and $X_2 \in \mathcal{X}_{p_2}$ of rank r_2. Let c^* denote the largest c such that X_1, X_2 admit a $(c, r_1 - c, r_2 - c)$-factorization. Then $c^* = r_1 + r_2 - r_+$, where $r_+ := rank([X_1, X_2])$.*

Proof. Our proof is nothing more than mapping the notions in our problem to the notions in vector space. Some mapping turned out to be non-trivial.

Define $V_t := span(X_t)$ for $t = 1, 2$. Then the first two mappings are trivial, i.e. $dim(V_t) = rank(X_t)$.

For mapping $dim(V_1 + V_2) = rank([X_1, X_2])$, we will prove $V_1 + V_2 = span([X_1, X_2])$ from two directions. For any vector $v \in V_1 + V_2$, it is a sum of two vectors, each from one V_t and thus can be expressed as a linear column sum [4] of X_t. Therefore v can be expressed a linear column sum of $[X_1, X_2]$ and thus in its span, which proves one direction $V_1 + V_2 \subseteq span([X_1, X_2])$. On the other hand, any $u \in span([X_1, X_2])$ is a linear column sum of $[X_1, X_2]$, which can always be split into two parts: the linear column sum of X_1 and that of X_2. Each split term is in one $span(X_t) = V_t$, implying $u \in V_1 + V_2$ and thus $span([X_1, X_2]) \subseteq V_1 + V_2$.

It remains to prove $c^* = dim(V_1 \cap V_2)$. For convenience, denote $k^*_t = r_t - c^*$ and the (c^*, k^*_1, k^*_2)-factorization admitted by X_1, X_2 as $X_t = \bar{D}^* P_t + \tilde{D}^*_t Q_t$, where $\bar{D}^* \in \mathbb{R}^{n \times c^*}$ and $\tilde{D}^*_t \in \mathbb{R}^{n \times k^*_t}$. Since this is full-rank factorization, it is easy to verify that $span(X_t) = span([\bar{D}^*, \tilde{D}^*_t])$. Note by definition $c^* = rank(\bar{D}^*)$, and thus to prove $c^* = dim(V_1 \cap V_2)$, it suffices to prove $span(\bar{D}^*) = V_1 \cap V_2$.

Let us first prove $span(\bar{D}^*) \subseteq V_1 \cap V_2$. For any $v \in span(\bar{D}_*)$, since $span(\bar{D}_*) \subseteq span(X_t)$ for $t = 1, 2$, it follows $v \in span(X_1) \cap span(X_2) = V_1 \cap V_2$. Next, we prove $V_1 \cap V_2 \subseteq span(\bar{D}^*)$ by contradiction. Suppose there exists a $u \in span(X_1) \cap span(X_2)$ but $u \notin span(\bar{D}^*)$. This means $\bar{D}^*$ is not sufficient to express u. Without loss of generality, let us assume we have to add one more column q to $\bar{D}^*$ for expressing u. Define $\bar{D}^*_+ := [\bar{D}^*, q]$ and we have $rank(\bar{D}^*_+) = c^* + 1$. Notice we can always construct q by a linear column sum of $[\tilde{D}_1, \tilde{D}_2]$, because u stays in $span(X_1) \cap span(X_2)$. As a result, we have $span(\bar{D}^*_+) \subseteq span(X_t)$ for $t = 1, 2$, which implies X_1, X_2 at least admit a $(c^* + 1, k^*_1 - 1, k^*_2 - 1)$-factorization. This contradicts with our premise that c^* is the largest number of an admitted factorization, and thus $span(X_1) \cap span(X_2) \subseteq span(\bar{D}^*)$. Combining two directions we conclude $span(\bar{D}^*) = V_1 \cap V_2$ and thus $c^* = rank(span(\bar{D}_*)) = dim(V_1 \cap V_2)$.

Now, all notions in our problem are mapped to those in the vector space. Applying Lemma 10 proves the corollary. □

Implementation and Simulation

In this section we present an algorithm based on the pCMF model and simulate its performance.

A simple pCMF-based algorithm

Let $W_t \in \mathbb{R}^{n \times p_t}$ denote the binary *mask* of sample υ_t such that $W_t(i, j) = 1$ for all $(i, j) \in \upsilon_t$, and $W_t(i, j) = 0$ for all $(i, j) \notin \upsilon_t$. Pack $\hat{\Gamma} := \{P_1, P_2, Q_1, Q_2\}$ and let $A \circ B$

[4] A linear column sum of X is a sum of the columns of X

denotes the Hadarmat product of matrices A and B. We proposed a simple pCMF algorithm that finds $\hat{\Theta}, \hat{\Gamma}$ to minimize the following objective function

$$f(\hat{\Theta}, \hat{\Gamma}) = \sum_t W_t \circ ||X_t - \bar{D}P_t - \tilde{D}_t Q_t||_F^2 + \lambda \left(||\bar{D}||_F^2 + \sum_t ||\tilde{D}_t||_F^2 + ||P_t||_F^2 + ||Q_t||_F^2 \right), \quad (30)$$

where $||\cdot||_F$ denotes the F-norm and λ is the regularization coefficient. To find the optimum, we applied the same alternate optimization method in (Singh and Gordon 2008).

Experimental data and protocol

To perform simulation, we construct a synthetic data set as follows: first, set $n = 200$, $p_1 = 300$, $p_2 = 500$, $\lambda = 0.01$ and set both matrix ranks as 20 so that $c+k_1 = c+k_2 = 20$, where c, k_1, k_2 will vary in experiment. Let E_n be an identity matrix of dimension n. Then the (ground truth) factors are constructed by selecting from E_n the first k_1 columns to form $\tilde{D}_1$, the last k_2 columns to form $\tilde{D}_2$, and the $k_1 + 1$ to $k_1 + c$ columns to form $\bar{D}$. It is clear that different factors span different subspaces. Then we randomly generated A_t of dimension $c+K_t$-by-p_t as loadings [5]. Now, the two matrices are generated by $X_t = [\bar{D}, \tilde{D}_t]A_t$ for $t = 1, 2$, so that their ranks are both 20 and they admit a (c, k_1, k_2)-factorization.

Three methods will be examined in experiment, including our proposed pCMF algorithm, the classic CMF method and independent matrix factorization (IMF) that factorizes and recovers each matrix independently. For both IMF and CMF, the size of their estimated factors (either independent or shared) are set as n-by-20, and for pCMF, the size of estimates of $\bar{D}, \tilde{D}_t$ are set as n-by-c and n-by-k_t respectively.

During experiment, we varied c (and consequently k_1, k_2) from 0 to 18. At each setting, we randomly chose 10% of each matrix's entries as a sample, i.e. $|\upsilon_t|/|\mathcal{I}_{p_t}| = 0.1$, and used the estimated factors and loadings to recover the rest entries and reported the recovery error in terms of rooted means squared error. At each setting of c, we repeated the random choice of sample for 20 times and reported the averaged error and the corresponding deviation in Figure 1.

Observations

From Figure 1 we see that CMF performed worse than IMF when c is small, i.e. when X_1 and X_2 are actually sharing very few part of their factors. When c increased, CMF improved (since its assumption was better fulfilled) and started to outperform IMF. This is the well-known negative transfer effect observed in the literature. On the other hand, our proposed pCMF method consistently outperformed IMF under different values of c, showing that its effectiveness and robustness in addressing negative transfer.

Another observation is that both CMF and pCMF gained smaller variance compared with IMF at all values of c. This is probably because their (partially) shared factor resulted in less parameters for estimate, and the shared factor was estimated with a larger sample (of size $|\upsilon_+|$).

[5] The entries of A_t are drawn from a standard normal distribution and, as we verified, a typical A_t has full row rank

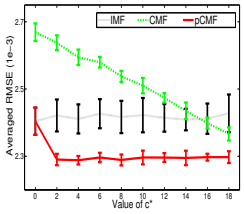

Figure 1: Averaged RMSE of two matrices versus c^*. Results are averaged over 20 random trials and the mean error bars with standard deviations are presented.

Conclusion

In this paper we formalized a partial collective matrix factorization model and imported the PAC framework from computational learning theory to analyze its error bound. We tackled the additional technical challenges that matrix entries are not sampled with replacement, neither could we decompose the sample error into the errors of each individual entry. We re-cast a coupled two-matrix recovery problem into three independent latent-matrix recovery problems under the PAC framework, and applied single matrix analysis on each to derive the final error bound. Under mild conditions, we also identified the optimal choice of parameters by mapping our problem into the problem of relating subsapce dimensions in a vector space, which provided guidance for algorithm design.

Our partial factorization model not only enjoyed a theoretical guarantee, but also motivated a simple algorithm that, as showed in our simulation, can avoid negative transfer and consistently outperformed both CMF and IMF methods.

References

Agarwal, D.; Chen, B.-C.; and Long, B. 2011. Localized factor models for multi-context recommendation. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 609–617. ACM.

Bouchard, G.; Yin, D.; and Guo, S. 2013. Convex collective matrix factorization. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, 144–152.

Candès, E. J., and Recht, B. 2009. Exact matrix completion via convex optimization. *Foundations of Computational mathematics* 9(6):717–772.

Ding, G.; Guo, Y.; and Zhou, J. 2014. Collective matrix factorization hashing for multimodal data. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2083–2090. IEEE.

Gunasekar, S.; Yamada, M.; Yin, D.; and Chang, Y. 2015. Consistent collective matrix completion under joint low rank structure. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 306–314.

Hofmann, D. C. T. 2001. The missing link-a probabilistic model of document content and hypertext connectivity. In *Proceedings of the 2000 Conference on Advances in Neural Information Processing Systems. The MIT Press*, 430–436.

Kearns, M. J., and Vazirani, U. V. 1994. *An introduction to computational learning theory*. MIT press.

Klami, A.; Bouchard, G.; Tripathi, A.; et al. 2014. Group-sparse embeddings in collective matrix factorization. In *Proceedings of International Conference on Learning Representations (ICLR) 2014*.

Lippert, C.; Weber, S. H.; Huang, Y.; Tresp, V.; Schubert, M.; and Kriegel, H.-P. 2008. Relation prediction in multi-relational domains using matrix factorization. In *NIPS Workshop: Structured Input-Structured Output*.

Maurer, A., and Pontil, M. 2010. K-dimensional coding schemes in hilbert spaces. *Information Theory, IEEE Transactions on* 56(11):5839–5846.

Prasad, A. 2010. Counting subspaces of a finite vector space1. *Resonance* 15(11):977–987.

Rudi, A.; Canas, G. D.; and Rosasco, L. 2013. On the sample complexity of subspace learning. In *Advances in Neural Information Processing Systems*, 2067–2075.

Singh, A. P., and Gordon, G. J. 2008. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 650–658. ACM.

Singh, A., and Gordon, G. 2010. A bayesian matrix factorization model for relational data. In *UAI*.

Srebro, N.; Alon, N.; and Jaakkola, T. S. 2004. Generalization error bounds for collaborative prediction with low-rank matrices. In *Advances In Neural Information Processing Systems*, 1321–1328.

Tan, M.; Tsang, I. W.; Wang, L.; Vandereycken, B.; and Pan, S. J. 2014. Riemannian pursuit for big matrix recovery. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 1539–1547.

Wade, W. 2003. *An Introduction to Analysis*. Pearson Prentice Hall.

Warner, S. 1965. *Modern algebra*. Number v. 1 in Prentice-Hall mathematics series. Prentice-Hall.

Yang, L.; Jing, L.; and Ng, M. K. 2015. Robust and non-negative collective matrix factorization for text-to-image transfer learning. *IEEE Transactions on Image Processing* 24(12):4701–4714.

Yu, K.; Yu, S.; and Tresp, V. 2005. Multi-label informed latent semantic indexing. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 258–265. ACM.

Zhang, Y.; Cao, B.; and Yeung, D.-Y. 2010. Multi-domain collaborative filtering. In *UAI*.

Zhou, J.; Wang, F.; Hu, J.; and Ye, J. 2014. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 135–144. ACM.