Towards a Theoretical Understanding of Negative Transfer in Collective Matrix Factorization

Chao Lan
EECS Department
University of Kansas
Lawrence, KS 66045
clan@ittc.ku.edu

Jianxin Wang
School of Information Science and Engineering
Central South University
Changsha, China
jxwang@mail.csu.edu.cn

Jun Huan
EECS Department
University of Kansas
Lawrence, KS 66045
jhuan@ittc.ku.edu

Abstract

Collective matrix factorization (CMF) is a popular technique to improve the overall factorization quality of multiple matrices presuming they share the same latent factor. However, it suffers from performance degeneration when this assumption fails, an effect called *negative transfer (n.t.)*. Although the effect is widely admitted, its theoretical nature remains a mystery to date.

This paper presents a first theoretical understanding of $n.t.$ in theory. Under the statistical minimax framework, we derive lower bounds for the CMF estimator and gain two insights. First, the $n.t.$ effect can be explained as the rise of a bias term in the standard lower bound, which depends only on the structure of factor space but neither the estimator nor samples. Second, the $n.t.$ effect can be explained as the rise of an d_{th}-root function on the learning rate, where d is the dimension of a Grassmannian containing the subspaces spanned by latent factors. These discoveries are also supported in simulation, and suggest $n.t.$ may be more effectively addressed via model construction other than model selection.

1 INTRODUCTION

Collective matrix factorization (CMF) is a popular technique to factorize multiple matrices in hope of improving their overall factorization quality (e.g. [7, 23, 17, 11, 16, 2, 12, 3, 21, 26]). The key assumption of CMF is that all matrices share the same low-rank factor, under which its estimator proves to be consistent [5]. However, when this assumption fails, CMF is known to suffer from performance degeneration – an effect called *negative transfer*. Several algorithmic solutions have been proposed, which alternatively assume different matrix factors are drawn from the same distribution [24, 1] or partially shared [10].

Although negative transfer has been long accused for causing the performance degeneration, the theoretical understanding on its nature appears surprisingly scarce, i.e. no study was done to justify its existence or how it may hurt CMF. *What can we say about negative transfer in theory?* This is the question we aim to address in the paper.

Our investigation is performed under the mini-max framework in statistical decision theory. We first cast CMF into this framework and design a collective hypothesis testing problem that captures the negative transfer effect. By reducing the CMF estimation problem to this testing problem, we manage to derive a lower bound of the CMF estimator, through which a new bias term is discovered that worsens the standard bound. In particular, the bias only depends on the structure of the factor space, but neither the choice of estimator nor training samples. This suggests negative transfer is an intrinsic difficulty of learning, which may only be resolved at the model construction phase but neither model selection nor data collection. This is also supported from another observation that negative transfer down-weights the contribution of estimation accuracy in the lower bound.

For better interpretability, we further refine the lower bound by capturing more problem characteristics. In particular, we derive a learning rate of $\Omega(1/|\omega|^{\frac{1}{d}})$, where $\vec{\omega}$ is the index set of all matrix observations and d is the dimension of a Grassmannian containing the subspaces spanned by latent factors. Pessimistically, this rate is d_{th}-root slower than the standard rate $\Omega(1/|\omega|)$ where negative transfer does not exist. This discovery is also supported in our simulation.

The rest of this paper is organized as follows: the notations are introduced in section two; our primary lower bound is presented in section three, and the refined bound is presented in section four; proofs and remarks are given in section five, followed by simulation in section six and conclusions in section seven.

2 PRELIMINARIES

In this section we introduce the major notations, concepts and assumptions used in analysis. For the ease of presentation, we focus on two matrix factorization, but all discussions are readily generalizable.

Matrix Notations. For a matrix M, let M_{ij} be its entry at row i and column j, let $[M]$ be its column space, $||M||$ be its Frobenius norm [1] and M^T be its transpose. Given two matrices M, M' of the same column size, let $\vec{M} = [M, M']$ be their column concatenation. Let I be an identify matrix properly sized by the context.

Sets. Let $\mathbb{M}_k^{n\cdot}$ be a set of rank-k matrices with row dimension n and arbitrary column dimension, and $\mathbb{M}_k^{n,p} \subseteq \mathbb{M}_k^{n\cdot}$ be its subset with column dimension p. Let $\mathbb{G}_k^n$ be the Grassmannian defined as the set of k-dimensional subspaces in $\mathbb{R}^n$ (a metric will be equipped later). Note each element in $\mathbb{G}_k^n$ is a subspace. Let $\mathbb{S}_k^n$ be the set of orthonormal matrices in $\mathbb{R}^{k\times n}$.

Metrics. Let d be the metric on $\mathbb{M}_k^{n,p}$ such that

$$d(M, M') = ||M - M'||, \tag{1}$$

for all $M, M' \in \mathbb{M}_k^{n,p}$. It will be used for any choice of p. Let ρ be the metric on $\mathbb{G}_k^n$ such that for any $\mathcal{G}, \mathcal{G}' \in \mathbb{G}_k^n$,

$$\rho(\mathcal{G}, \mathcal{G}') = ||\mathrm{P}_{\mathcal{G}} - \mathrm{P}_{\mathcal{G}'}||, \tag{2}$$

where $\mathrm{P}_{\mathcal{G}}$ is the orthogonal projection matrix onto $\mathcal{G}$. It is defined as $\mathrm{P}_{\mathcal{G}} = DD^T$ for any basis $D \in \mathbb{S}_k^n$ of $\mathcal{G}$. See [4, Section 2.5] for more explanations.

Factorization Model. We focus on full-rank matrix factorization, which is typically assumed in CMF (e.g. [17, 5]). The factorization model is generally denoted as $M = DA$, where D is called the *factor* and A is called the *loading*. For the ease of presentation, we assume $D \in \mathbb{S}_k^n$ but all discussions are generalizable. (See Remark 11 for a justification.) In general the loading will not be specified in analysis, as long as it is a properly sized full rank matrix. All matrices are assumed bounded.

The Shared-Factor Assumption of CMF. The CMF assumption can be stated as: *any input $M, M' \in \mathbb{M}_k^{n\cdot}$ admit factorization $M = DA$ and $M' = D'A'$ such that $[D] = [D']$*. Note although CMF assumes $D = D'$, in essence it only requires $[D] = [D']$. Also note $[D] \in \mathbb{G}_k^n$.

Probability Notations. We mainly use two styles of probability notations with different focuses: 1) notation $\Pr\{\cdot\}$ focuses on the uncertainty over random samples, e.g. in section 3 where we prove the mini-max bound randomized over samples; 2) notation $\mathbb{P}$ focuses on treating the probability as a subject of interest, e.g. in section 4 where we pick a finite set of probabilities for testing which one generates the random sample.

Sampling Model. In many applications M is not fully observed (e.g. matrix completion [17]). Let ω be the index set of observed entries and M_ω be the input matrix. Assume ω is randomly sampled. Given a concatenated matrix $\vec{M}$, we use $\vec{\omega}$ to denote its index set of observations (induced from the observations of each matrix).

Generative Model. The matrix generative model is needed for refining the mini-max bound, but not for modeling negative transfer. In our analysis, only one generative model needs to be defined on the concatenated matrix $\vec{M}$.

Suppose $\vec{M} \in \mathbb{M}_k^{n,\vec{p}}$. Let $\mathcal{P}$ be a set of probabilities defined on $\mathbb{M}_k^{n,\vec{p}}$ such that each $\mathbb{P} \in \mathcal{P}$ is a matrix-variate normal distribution (e.g. [6, Chapter 2]),

$$\mathbb{P}_{\bar{M}}(\vec{M}) = \mathcal{N}(\bar{M}, \sigma^2\mathrm{I}, \sigma^2\mathrm{I}'), \tag{3}$$

with mean matrix $\bar{M}$ and covariance matrices $\sigma^2\mathrm{I}$ (among rows) and $\sigma^2\mathrm{I}'$ (among columns). It follows

$$\mathbb{P}_{\bar{M}}(\vec{M}) = \prod\nolimits_{(i,j)} \tilde{\mathbb{P}}_{\bar{M}}(\vec{M}_{ij}), \tag{4}$$

where $\tilde{\mathbb{P}}_{\bar{M}}(\vec{M}_{ij}) = \tilde{\mathcal{N}}(\bar{M}_{ij}, \sigma^2)$ is a univariate normal distribution and the product is taken over all indices of $\vec{M}$. We note in passing $\mathbb{P}_{\bar{M}}$ is similar to the probabilistic matrix factorization model assumed in [15, Equation 1]. Define

$$\mathbb{P}_{\bar{M}}(\vec{M}_{\vec{\omega}}) = \prod\nolimits_{(i,j)\in\vec{\omega}} \tilde{\mathbb{P}}_{\bar{M}}(\vec{M}_{ij}). \tag{5}$$

Mapping. Since each M admits a unique $[D]$ in factorization $M{=}DA$ (Remark 12), we have a mapping $\theta : \mathbb{M}_k^{n\cdot} \to \mathbb{G}_k^n$ such that $\theta(M) = [D]$. This is the mapping from a rank-k matrix with n rows to its column space in $\mathbb{R}^n$. Note θ applies to any column dimension.

CMF Estimator. Recall $\vec{M}_{\vec{\omega}}$ is the random observation of two matrices. Define CMF estimator as $\hat{\theta} : \{\vec{M}_{\vec{\omega}}\} \to \mathbb{G}_k^n$. Note it realizes the shared-factor assumption by mapping two matrix observations into a single subspace. Write $\hat{\theta}_{\vec{\omega}}$ for $\hat{\theta}(\vec{M}_{\vec{\omega}})$. The quality of $\hat{\theta}$ is evaluated by

$$\ell_{\vec{\omega}}(\hat{\theta}|\vec{M}) = \frac{1}{2}\left[\rho(\hat{\theta}_{\vec{\omega}}, \theta(M)) + \rho(\hat{\theta}_{\vec{\omega}}, \theta(M'))\right]. \tag{6}$$

Define the maximum risk of any CMF estimator as

$$\mathfrak{M}(\hat{\theta}) = \sup_{\vec{M}} \mathbb{E}_{\vec{\omega}} \ell_{\vec{\omega}}(\hat{\theta}|\vec{M}), \tag{7}$$

where expectation $\mathbb{E}_\omega$ is taken over the randomness of $\vec{M}_{\vec{\omega}}$.

It was noted when the shared-factor assumption is satisfied, CMF is equivalent to factorizing on a single matrix $\vec{M}$ [10].

Packing. This notion is used to define the following hypothesis testing problem and widely used in proof. For any

[1] This notation should not cause confusion since we only consider Frobenius norm in this paper.

set $\mathcal{X}$ equipped with a metric $\rho_{\mathcal{X}}$, let $\{x_v\}_{v\in\mathcal{V}}$ be an arbitrary subset of $\mathcal{X}$ indexed by a set $\mathcal{V}$. For any $\delta > 0$, we say this subset is a δ-packing of $\mathcal{X}$ with respect to $\rho_{\mathcal{X}}$ if $\rho_{\mathcal{X}}(x_v, x_{v'}) \geq \delta$ whenever $v \neq v'$.

Collective Hypothesis Testing. To lower bound $\mathfrak{M}(\hat{\theta})$, we employ the classic estimation-to-testing reduction method (e.g. [22]). To capture the negative transfer effect, we additionally design a *collective hypothesis testing* problem.

Let $\{\mathcal{G}_v\}_{v\in\mathcal{V}}$ be a 2δ-packing of $\mathbb{G}_k^n$ indexed by a finite set $\mathcal{V}$, and V, V' be two random variables taking values $v, v' \in \mathcal{V}$ respectively. Our testing problem is stated as follows:

Step 1: choose V, V' (with replacement) from $\mathcal{V}$ independently and uniformly at random.

Step 2: conditioned on $(V = v, V' = v')$, randomly choose $(D, D') \in \mathbb{S}_k^n \times \mathbb{S}_k^n$ satisfying $([D], [D']) = (\mathcal{G}_v, \mathcal{G}_{v'})$.

Step 3: generate $(M, M') = (DA, D'A')$ with some random A, A'; then generate observation $\vec{M}_{\vec{\omega}}$ from $\vec{M}$.

Step 4: apply a collective testing function $\hat{V} : \{\vec{M}_{\vec{\omega}}\} \to \mathcal{V}$ defined as [2]

$$\hat{V}(\vec{M}_{\vec{\omega}}) := \arg\min_{v\in\mathcal{V}} \rho(\hat{\theta}(\vec{M}_{\vec{\omega}}), \mathcal{G}_v). \tag{8}$$

To our knowledge, the collective hypothesis testing problem, albeit simple, is the first attempt to theoretically model the negative transfer effect in CMF. It also allows incorporation of richer information such as prior distribution, and can be applied to other problems besides CMF, as will be exemplified in later discussions.

3 A PRIMARY LOWER BOUND

The following proposition presents a primary lower bound, which reveals our main idea and insights.

Proposition 1. *Suppose $\mathbb{G}_k^n$ admits a 2δ-packing indexed by a finite set $\mathcal{V}$, and V is a uniform random variable on $\mathcal{V}$. Then, any CMF estimator $\hat{\theta}$ satisfies*

$$\mathfrak{M}(\hat{\theta}) \geq \frac{\delta}{2} \cdot \left(C_\delta + \frac{1}{|\mathcal{V}|} \Pr\{\hat{V}(\vec{M}_{\vec{\omega}}) \neq V\} \right), \tag{9}$$

where $C_\delta = 1 - |\mathcal{V}|^{-1}$ and the probability is defined over the random choice of V and $\vec{M}_{\vec{\omega}}$ [3].

Our main idea of proving the proposition is to first reduce the estimation problem into the collective hypothesis testing problem. Then, if two matrices are generated from different subspaces (i.e. $\theta(M) \neq \theta(M')$), $\hat{V}$ is guaranteed to make mistake on at least one matrix by a geometrical argument in Figure 1. This inevitable mistake gives rise to bias C_δ in the lower bound, which does not depend on the choice of estimator nor observations. On the other hand, the testing error $\Pr\{\hat{V}(\vec{M}_{\vec{\omega}}) \neq V\}$ is obtained by standard arguments, conditioned on the case when two matrices are indeed generated from the same subspace.

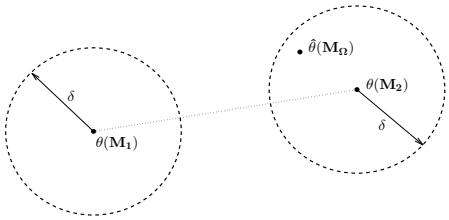

Figure 1: When two matrices are generated from different subspaces, no CMF estimator $\hat{\theta}$ can simultaneously fall into the two δ-balls centered at $\theta(M_2)$ and $\theta(M_2)$ respectively. As a result, the testing function $\hat{V}$ is guaranteed to make a mistake on at least one matrix.

Next, we discuss the implications of Proposition 1. For comparison, consider the case when negative transfer does not exist (i.e. $V = V'$). In this case, by Remark 13 we have

$$\mathfrak{M}(\hat{\theta}) \geq \delta \cdot \Pr\{\hat{V}(\vec{M}_{\vec{\omega}}) \neq V\}. \tag{10}$$

Comparing the lower bounds in (9) and (10), we see negative transfer has caused two changes: 1) introduce a bias term C_δ; 2) down-weight the testing error by $1/(2|\mathcal{V}|)$. In particular, the bias term depends merely on the structure of $\mathbb{G}_k^n$, pessimistically suggesting that CMF may never succeed in a mini-max sense, disregarding the choice of estimator $\hat{\theta}$ or observation $\vec{M}_{\vec{\omega}}$. Combining both aspects, it seems finding a good factor space may be more important than finding a good estimator or sample for mitigating the negative transfer effect.

When negative transfer does not exist, we may obtain another indirect implication by comparing CMF with independent matrix factorization (IMF) [4]. To elaborate the latter, let $\hat{\theta}_s : \{M_\omega\} \to \mathbb{G}_k^n$ be an IMF estimator and define its maximum risk as

$$\mathfrak{M}_s(\hat{\theta}_s) = \sup_M \mathbb{E}\, \rho(\hat{\theta}_s(M_\omega), \theta(M)), \tag{11}$$

where the expectation is taken over the randomness of ω. Let $\hat{V}_s : \{M_\omega\} \to \mathcal{V}$ be any testing function on a single matrix, which possibly induces $\hat{V}$. By standard mini-max arguments it is easy to verify that

$$\mathfrak{M}_s(\hat{\theta}_s) \geq \delta \cdot \Pr\{\hat{V}_s(M_\omega) \neq V\}. \tag{12}$$

A comparison between the lower bounds in (10) and (12) suggests CMF estimator performs *no worse* than IMF estimator on at least one matrix. The reason is $\hat{V}(\vec{M}_{\vec{\omega}}) \neq V$ implies inclusively either $\hat{V}_s(M_\omega) \neq V$ or $\hat{V}_s(M'_{\omega'}) \neq V$.

[2] Note $\vec{M}_{\vec{\omega}}$ depends on v, v' in Steps 3 and 4.

[3] Both V and $\hat{V}(\vec{M}_{\vec{\omega}})$ are random variables on $\mathcal{V}$.

[4] This is the technique that separately factorizes each matrix based on its own observations.

(Otherwise, $\hat{V}$ would not make the mistake if, say, it is simply defined as the random choice of one $\hat{V}_s$.) Taking M_w for instance, we thus have

$$\Pr\{\hat{V}(\vec{M}_{\vec{\omega}}) \neq V\} \leq \Pr\{\hat{V}_s(M_\omega) \neq V\}. \quad (13)$$

Further detailing (13) is beyond the scope of this paper. Nevertheless, it is not difficult to conjecture the inequality holds typically when M has insufficient observation (thus CMF improves over IMF on at least one matrix) and the equality holds otherwise. It should be noted even without negative transfer, our lower bound does not suggest CMF always improves over IMF. We note this is neither concluded from the upper bound analysis of CMF assuming no negative transfer [5].

Next we present two extensions of the proposition.

3.1 A BOUND WITH PRIOR

In [16], authors assumed a prior distribution over the factor space. A natural question for us is how such intrinsic prior may affect the negative transfer effect. For simplicity assume all sets are measurable.

Let ν be a probability measure defined on the factor space $\mathbb{S}_k^n$, as assumed in [16]. It naturally induces a probability measure μ over $\mathbb{G}_k^n$ such that for any $\mathcal{G} \in \mathbb{G}_k^n$,

$$\mu(\mathcal{G}) := \int_{[D]=\mathcal{G}} 1 \, d\nu(D). \quad (14)$$

Define $\tilde{\mu} := \mu/N$ as a normalized probability measure with proper choice of N. We can replace *Step 1* in the collective hypothesis testing problem with

Step 1^**: choose both* V, V' *(with replacement) by* $\tilde{\mu}(\mathcal{G}_V)$.

By the same arguments for Proposition 1, and now

$$\Pr\{V = V'\} = \sum\nolimits_{v \in \mathcal{V}} \tilde{\mu}^2(\mathcal{G}_v), \quad (15)$$

it is easy to verify the following result.

Corollary 2. *Suppose* $\mathbb{G}_k^n$ *admits a* 2δ*-packing indexed by a finite set* $\mathcal{V}$*. Let* V *be a uniform random variable on* $\mathcal{V}$*. Replace Step* 1 *in collective hypothesis testing with Step* 1^**. Then, any CMF estimator* $\hat{\theta}$ *satisfies*

$$\mathfrak{M}(\hat{\theta}) \geq \frac{\delta}{2} \cdot \left(\tilde{C}_\delta + \tilde{N}_\delta \Pr\{\hat{V}(\vec{M}_{\vec{\omega}}) \neq V\} \right), \quad (16)$$

where $\tilde{N}_\delta = \sum_{v \in \mathcal{V}} \tilde{\mu}^2(\mathcal{G}_v)$*,* $\tilde{C}_\delta = 1 - \tilde{N}_\delta$*, and the probability is defined over the random choice of* $\vec{M}_{\vec{\omega}}$ *and* V*.*

The implication of Corollary 2 is clear: since $\tilde{N}_\delta$ reaches its minimum when $\tilde{\mu}(\mathcal{G}_V)$ is the same for all choices of V (by Chebyshev's sum inequality), resulting in the maximum bias $\tilde{C}_\delta$, we see CMF suffers most negative transfer when nature chooses V uniformly.

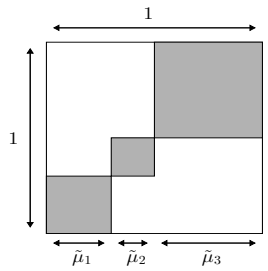

Figure 2: A packing of three points, each indexing a gray square in the figure with its prior $\mu_i := \mu(\mathcal{G}_i)$, $i = 1, 2, 3$. The total area of three squares equals to $\tilde{N}_\delta$. Clearly, this area reaches its maximum value 1 when $\mu_i = 1$ for any i. This corresponds to the most concentrated prior μ.

On the other hand, a simple geometric argument in Figure 2 shows the more concentrated $\tilde{\mu}$ nature uses, the less negative transfer CMF suffers.

3.2 A BOUND FOR MATRIX RECOVERY

A major use of CMF is for recovering the missing values of incomplete matrices (e.g. [17]). This section presents a technique to extend Proposition 1 for the recovery task under mild conditions. The main strategy is to convert the recovery error back to $\rho(\hat{\mathcal{G}}, \mathcal{G})$ using the following variant of [18, Theorem 2.3].

Lemma 3. *Let* $\mathcal{G}, \mathcal{G}'$ *respectively be the column spaces of any* $M, M' \in \mathbb{M}_k^{n,p}$*. Let* $s_k(M)$ *be the smallest non-zero singular value of* M*. Then*

$$\rho(\mathcal{G}, \mathcal{G}') \leq \sqrt{2}||M - M'||/s_k(M). \quad (17)$$

For simplicity, we focus on a set $\tilde{\mathbb{M}}_k^{n\cdot} \subseteq \mathbb{M}_k^{n\cdot}$ whose matrices have their smallest non-zero singular values bounded away from zero. A similar assumption was made for matrix recovery in [9, Theorem I.2.].

Let $\tilde{\mathbb{G}}_k^n \subseteq \mathbb{G}_k^n$ be the set induced from $\tilde{\mathbb{M}}_k^{n\cdot}$ such that for every $\mathcal{G} \in \tilde{\mathbb{G}}_k^n$ there is an $M \in \tilde{\mathbb{M}}_k^{n\cdot}$ satisfying $\theta(M) = \mathcal{G}$. For a matrix M and a factor $\hat{D}$ estimated from its observation M_ω, define the recovery error as

$$er_M(\hat{D}) = \min_A ||M - \hat{D}A||^2. \quad (18)$$

This is similar to the reconstructive error in [13, page 1]. Define the recovery loss as

$$\ell_{\mathfrak{r}|\vec{\omega}}(\hat{\theta}|\vec{M}) = \frac{1}{2}\left[er_M(\hat{\theta}(\vec{M}_{\vec{\omega}})) + er_{M'}(\hat{\theta}(\vec{M}_{\vec{\omega}})) \right], \quad (19)$$

and the maximum risk of any CMF estimator as

$$\tilde{\mathfrak{M}}_{\mathfrak{r}}(\hat{\theta}) = \sup_{\vec{M} \in \tilde{\mathbb{M}}_k^{n\cdot} \times \tilde{\mathbb{M}}_k^{n\cdot}} \mathbb{E}_{\vec{\omega}} \ell_{\mathfrak{r}|\vec{\omega}}(\hat{\theta}|\vec{M}). \quad (20)$$

Our recovery bound is stated as follows.

Corollary 4. *Given a $\tilde{\mathbb{M}}_k^{n\cdot}$ and its induced $\tilde{\mathbb{G}}_k^n$ that admits a 2δ-packing indexed by a finite set $\mathcal{V}$. Let V be a uniform random variable on $\mathcal{V}$. Then, there is a $c > 0$ (depending on $\tilde{\mathbb{M}}_k^{n\cdot}$) bounded away from zero such that every CMF estimator $\hat{\theta}$ satisfies*

$$\tilde{\mathfrak{M}}_{\mathfrak{r}}(\hat{\theta}) \geq \frac{\delta \cdot c}{2\sqrt{2}} \left(C_\delta + \frac{1}{|\mathcal{V}|} \Pr\{\hat{V}(\vec{M}_{\vec{\omega}}) \neq V\} \right), \quad (21)$$

where $C_\delta = 1 - |\mathcal{V}|^{-1}$ and the probability is defined over the random choice of $\vec{M}_{\vec{\omega}}$ and V.

Corollary 4 shows a recovery error bound that maintains the same order as the estimation error bound.

4 A FINER LOWER BOUND

In this section, we first introduce a few more machinery used to refine the lower bound in Proposition 1, and then present the refined bound.

Generalized Fano Method for CMF. The first piece of information is related to the generalized Fano method in [22, Lemma 3], which will be the starting point of our proof.

Let $I(;)$ denote the mutual information between two variables. The following lemma is our extension of the generalized Fano method for the CMF problem.

Lemma 5. *Let $\{M_v \in \mathbb{M}_k^{n\cdot}\}_{v\in\mathcal{V}} \subseteq \mathcal{P}$ be a collection of matrices indexed by $\mathcal{V}$ such that for any $v \neq v'$,*

$$\rho(\theta(M_v), \theta(M_{v'})) \geq 2\delta. \quad (22)$$

Further, suppose

$$I(V; \vec{M}_{\vec{\omega}}) \leq \beta, \quad (23)$$

where V is a uniform random variable on $\mathcal{V}$. Then

$$\max_{v,v'\in\mathcal{V}} \mathbb{E}_\omega \frac{1}{2} \left(d(\hat{\theta}, \theta(M_v)) + d(\hat{\theta}, \theta(M_{v'})) \right) \geq \frac{\delta}{2} \left(1 - \frac{\beta + \log 2}{|\mathcal{V}| \log |\mathcal{V}|} \right). \quad (24)$$

Comparing (24) with the standard bound [22, Lemma 3] [5]

$$\frac{\delta}{2} \left(1 - \frac{\beta + \log 2}{\log |\mathcal{V}|} \right), \quad (25)$$

our generalization introduces an additional $|\mathcal{V}|$ in the denominator, which significantly speeds up the growth of the lower bound as $|\mathcal{V}|$ increases. This coincides with with our discovery in Proposition 1, and provides a finer implication of the negative transfer effect. In result we retain the mutual information (instead of relaxing it to KL divergence) to facilitate later application.

A few things should be clarified about the lemma. First, it does not require M_v's to have the same column dimension. Second, $I(V; \vec{M}_{\vec{\omega}})$ is derived from $\Pr\{\hat{v}(\vec{M}_{\vec{\omega}}) \neq V\}$ in Proposition 1 and thus inherits the condition that two matrices share the same latent factor.

Packing Number. The second piece of information is related to the packing number on Grassmannian $\mathbb{G}_k^n$, which will be used to bound $|\mathcal{V}|$ in Lemma 5.

Let $M(\mathbb{G}_k^n, \rho, \delta)$ be the packing number on $\mathbb{G}_k^n$ with respect to metric ρ and radius δ. It is the largest size of $\mathcal{V}$ that indexes an admitted δ-packing on $\mathbb{G}_k^n$. Let $\tau(\mathbb{G}_k^n)$ and d be the diameter and dimension of $\mathbb{G}_k^n$, respectively. The following result is a variant of [19, Proposition 8].

Lemma 6. *There exist universal constants $c_1, c_2 > 0$ such that for any $\delta \in (0, \tau(\mathbb{G}_k^n)]$,*

$$(c_1 \tau(\mathbb{G}_k^n)/\delta)^d \leq M(\mathbb{G}_k^n, \rho, \delta) \leq (c_2 \tau(\mathbb{G}_k^n)/\delta)^d. \quad (26)$$

Mutual Information. The third piece of information is related to the mutual information $I(V; \vec{M}_{\vec{\omega}})$ appeared in Lemma 5, which will be used for deriving its upper bound.

Let $D_{k\ell}$ denote KL divergence. Recall the probability notation $\mathbb{P}$ introduced in section 2. A classic approach (e.g. [22, page 428]) to bound $I(V; \vec{M}_{\vec{\omega}})$ is by

$$I(V; \vec{M}_{\vec{\omega}}) \leq \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} D_{k\ell}(\mathbb{P}_v || \mathbb{P}_{v'}). \quad (27)$$

However, this does not directly apply to our setting since $\mathbb{P}_v$ is not easy to specify. The following technique is from [8, Equation 110], which addresses the problem.

Lemma 7. *Let $T(\vec{M}_{\vec{\omega}})$ be any side information. Then*

$$I(V; \vec{M}_{\vec{\omega}}) \leq I(V; \vec{M}_{\vec{\omega}} | T(\vec{M}_{\vec{\omega}})). \quad (28)$$

Write $\vec{T}$ for $T(\vec{M}_{\vec{\omega}})$. We notice

$$I(V; \vec{M}_{\vec{\omega}} | \vec{T}) \leq \frac{\sum_{v,v'} \mathbb{E} D_{k\ell}(\mathbb{P}_v(\cdot|\vec{T}) || \mathbb{P}_{v'}(\cdot|\vec{T}))}{|\mathcal{V}|^2}, \quad (29)$$

where expectation $\mathbb{E}$ is taken over the randomness of $\vec{T}$.

KL Divergence. The last piece of information is related to KL divergence, which is used to specify the bound in (29).

Recall the generative model introduced in section 2. For a matrix $\vec{M}$ and its observation index $\vec{\omega}$, let W_ω be a matrix of the same size as $\vec{M}$ such that $W_{ij|\vec{\omega}} = 1$ if $(i,j) \in \vec{\omega}$ and $W_{ij|\vec{\omega}} = 0$ otherwise. Let $\circ$ denote the Hadamard product between matrices. We remark the following result.

Lemma 8. *For any $\bar{M}, \bar{M}' \in \mathbb{M}_k^{n,\vec{p}}$ and $\vec{\omega}$,*

$$D_{k\ell}(\mathbb{P}_{\bar{M}|\vec{\omega}} || \mathbb{P}_{\bar{M}'|\vec{\omega}}) = \frac{1}{2\sigma^4} ||W_{\vec{\omega}} \circ (\bar{M} - \bar{M}')||^2. \quad (30)$$

[5] While [22] focused on probability set, we focus on matrix set to facilitate later application. Nevertheless, our extension is also applicable on probability set and will give result similar to (24).

4.1 THE FINER BOUND

Recall the factorization model $M = DA$ and A is randomly generated. Let $\Sigma_A = \mathbb{E}||A||^2$. Our finer bound is based on the setting of Proposition 1 and stated as follows.

Theorem 9. *Every CMF estimator $\hat{\theta}$ satisfies*

$$\mathfrak{M}(\hat{\theta}) \geq c \cdot \tau(\mathbb{G}_k^n)^{1-1/d}(|\vec{\omega}|\Sigma_A/\sigma^4)^{-1/d}, \tag{31}$$

where $c > 0$ depends on the nature of $\mathbb{G}_k^n$ and absorbs lower order terms.

The new lower bound can be interpreted as follows.

- $|\vec{\omega}|$: larger observation number $|\vec{\omega}|$ leads to smaller lower bound. This makes sense, as more observations improve the accuracy of testing. However, we see its impact is significantly restricted by d, resulting in a learning rate $\Omega(|\vec{\omega}|^{-1/d})$. Based on (25) and our arguments for the theorem, one can easily derive a learning rate without negative transfer as $\Omega(|\vec{\omega}|^{-1})$. Thus we see negative transfer significantly slows down learning. This shall not be too surprising, however, since in Lemma 5 the lower bound has already become linearly dependent on $|\mathcal{V}|$ (instead of logarithmically) due to the negative transfer effect.

- d: for simplicity, assume $\tau(\mathbb{G}_k^n)|\vec{\omega}|\Sigma_A/\sigma^4 \geq 1$. Then, larger d leads to larger lower bound. In particular, a very large d significantly weakens the impact of other parameters (except $\tau(\mathbb{G}_k^n)$) on the lower bound. This coincides with our discovery in Proposition 1, where the impact of estimation quality (and now its related parameters) is down-weighted.

  We notice the dimension $d = k(n-k)$ is quadratic to matrix rank k and reaches its maximum at $k = n/2$ (and thus the worst bound). It is unclear how to explain such role of k, but we have another consistent observation based mainly on combinatoric arguments: Assume $\mathbb{M}_k^{n\cdot}$ is defined on a finite field of order q (which is common in problems such as recommendation system). Then the number of its column spaces (thus factor spaces) is the q-binomial coefficient $\binom{n}{k}_q$ based on [14]. Clearly, the more subspaces $\mathbb{M}_k^{n\cdot}$ induces, the more difficult estimation/testing will be. In particular, we notice $\binom{n}{k}_q$ is also a quadratic-style function of k and reaches its maximum at $k = n/2$.

- $\tau(\mathbb{G}_k^n)$: larger diameter of $\mathbb{G}_k^n$ leads to larger lower bound. This makes sense, since a larger hypothesis set admits a larger packing (see Lemma 6), resulting in a more challenging testing problem. In addition, we see the impact of diameter is slightly restricted by the dimension d of $\mathbb{G}_k^n$. Specifically, a large diameter hurts more when the dimension is high.

- Σ_A and σ: we are not particularly interested in these two terms, but note in passing that larger Σ_A or smaller σ leads to smaller lower bound.

- c : this coefficient arises from the universal constants in Lemma 6 that depend on the nature of $\mathbb{G}_k^n$. Then it absorbs lower order terms through derivation, but this shall not affect the order of interested parameters.

5 PROOFS AND REMARKS

Proof of Proposition 1.

Write $\hat{\theta}$ for $\hat{\theta}(\vec{M}_{\vec{\omega}})$, $\vec{V}$ for (V, V') and $\vec{v}$ for (v, v'). Let notation $\vee$ denote the logical disjunction. Following standard mini-max arguments, we first have

$$\begin{aligned}
&\sup_{\vec{M}} \mathbb{E}_{\vec{\omega}} \left[\rho(\hat{\theta}, \theta(M)) + \rho(\hat{\theta}, \theta(M'))\right] \\
&\geq \sup_{\vec{M}} \mathbb{E}\left[\delta \, \mathbf{1}\{\rho(\hat{\theta}, \theta(M)) \geq \delta \vee \rho(\hat{\theta}, \theta(M')) \geq \delta\}\right] \\
&= \delta \cdot \sup_{\vec{M}} \Pr\{\rho(\hat{\theta}, \theta(M)) \geq \delta \vee \rho(\hat{\theta}, \theta(M')) \geq \delta\},
\end{aligned} \tag{32}$$

where the inequality is based on the fact that total distance is greater than δ if any one distance is greater than δ.

Reducing the above estimation problem into the collective hypothesis testing problem (with a 2δ-packing $\{\theta_v\}_{v\in\mathcal{V}}$), we have

$$\begin{aligned}
&\sup_{\vec{M}} \Pr\{\rho(\hat{\theta}, \theta(M)) \geq \delta \vee \rho(\hat{\theta}, \theta(M')) \geq \delta\} \\
&\geq \frac{1}{|\mathcal{V}|^2} \sum_{\vec{v}} \Pr\{\rho(\hat{\theta}, \theta_v) \geq \delta \vee \rho(\hat{\theta}, \theta_{v'}) \geq \delta \mid \vec{V} = \vec{v}\},
\end{aligned} \tag{33}$$

where the coefficient is based on the uniform sampling assumption on $\mathcal{V}$ so that $\Pr\{\vec{V} = \vec{v}\} = 1/|\mathcal{V}|^2$.

Now we introduce the negative transfer effect. Consider two cases $v = v'$ and $v \neq v'$. Clearly the second one captures the violation of the shared-factor assumption. Then

$$\begin{aligned}
&\Pr\{\rho(\hat{\theta}, \theta_v) \geq \delta \vee \rho(\hat{\theta}, \theta_{v'}) \geq \delta \mid \vec{V} = \vec{v}\} \\
&= \Pr\{\rho(\hat{\theta}, \theta_v) \geq \delta \vee \rho(\hat{\theta}, \theta_{v'}) \geq \delta \mid v \neq v', \vec{V} = \vec{v}\} \\
&\quad \cdot \Pr\{v \neq v' \mid \vec{V} = \vec{v}\} \\
&\quad + \Pr\{\rho(\hat{\theta}, \theta_v) \geq \delta \vee \rho(\hat{\theta}, \theta_{v'}) \geq \delta \mid v = v', \vec{V} = \vec{v}\} \\
&\quad \cdot \Pr\{v = v' \mid \vec{V} = \vec{v}\} \\
&= 1 \cdot \Pr\{v \neq v' \mid \vec{V} = \vec{v}\} \\
&\quad + \Pr\{\rho(\hat{\theta}, \theta_v) \geq \delta \vee \rho(\hat{\theta}, \theta_{v'}) \geq \delta \mid v = v', \vec{V} = \vec{v}\} \\
&\quad \cdot \Pr\{v = v' \mid \vec{V} = \vec{v}\},
\end{aligned} \tag{34}$$

where the second equality is based on the geometric argument illustrated in Figure 1, i.e. if $v \neq v'$, then no $\hat{\theta}$ can be

simultaneously δ-close to both θ_v and $\theta_{v'}$, which implies $\Pr\{\rho(\hat{\theta},\theta_v)\geq\delta\vee\rho(\hat{\theta},\theta_{v'})\geq\delta\}=1$.

Putting all above arguments together and in addition: 1) $\rho(\hat{\theta},\theta_{v_t})\geq\delta$ as implied by $\hat{V}(\vec{M}_{\vec{\omega}})\neq v_t$ (by the definition of $\hat{V}$); 2) average over all possible $\vec{v}$, we have

$$\begin{aligned}&\sup_{\vec{M}}\mathbb{E}\left[\rho(\hat{\theta},\theta(M))+\rho(\hat{\theta},\theta(M'))\right]\\&\geq\delta\left(\Pr\{V\neq V'\}+\frac{1}{|\mathcal{V}|}\Pr\{\hat{V}(\vec{M}_{\vec{\omega}})\neq V|V=V'\}\right).\end{aligned}\tag{35}$$

It remains to simplify the above lower bound. First, by uniform sampling $\Pr\{V\neq V'\}=1-|\mathcal{V}|^{-1}$. Second, $\Pr\{\hat{V}\neq V|V=V'\}=\Pr\{\hat{v}\neq V\}$, where the left side probability is over the randomness of both V,V', while the right side probability is merely over the randomness of V. Putting all together proves the proposition.

Proof of Corollary 4.

Recall $s_k(M)$ is the smallest non-zero singular value of matrix M. By our assumption $c=\inf_{M\in\tilde{\mathbb{M}}_k^{n\cdot}}s_k(M)$ is positive and bounded away from zero. Combining with Lemma 3, this implies any $M,M'\in\tilde{\mathbb{M}}_k^{n\cdot}$ satisfy

$$\rho(\mathcal{G},\mathcal{G}')\leq\sqrt{2}||M-M'||/c.\tag{36}$$

Writing $\hat{\theta}:=\hat{\theta}(\vec{M}_\omega)$, this further implies

$$er_M(\hat{\theta})\geq c\cdot\rho(\hat{\theta},\theta(M))/\sqrt{2}.\tag{37}$$

Hence over all $\vec{M}\in\tilde{\mathbb{M}}_k^{n\cdot}\times\tilde{\mathbb{M}}_k^{n\cdot}$, we have

$$\begin{aligned}&\sup_{\vec{M}}\mathbb{E}_{\vec{\omega}}\left[er_M(\hat{\theta})+er_{M'}(\hat{\theta})\right]\\&\geq\frac{c}{\sqrt{2}}\sup_{\vec{M}}\mathbb{E}_{\vec{\omega}}\left[\rho(\hat{\theta},\theta(M))+\rho(\hat{\theta},\theta(M'))\right].\end{aligned}\tag{38}$$

Applying Proposition 1 yields the corollary.

Proof of Lemma 5.

The proof is similar to [22, Lemma 3], with the main difference that we study a joint estimation problem (instead of a single one) and apply our own reduction technique.

By assumption $\{\phi(\mathbb{P}_v)\}_{v\in\mathcal{V}}$ is a 2δ-packing on $\mathbb{G}_k^n$. Applying the arguments in Proposition 1 gives a lower bound

$$\frac{\delta}{2}\left(C_\alpha+\frac{1}{|\mathcal{V}|}\Pr\{\hat{V}(\vec{M}_{\vec{\omega}})\neq V\}\right),\tag{39}$$

where $C_\alpha=1-|\mathcal{V}|^{-1}$.

Further, following the same arguments in the generalized Fano method [22, Lemma 3] (in particular, the Fano's inequality and data processing inequality), we have

$$\Pr\{\hat{V}(\vec{M}_{\vec{\omega}})\neq V\}\geq1-\frac{I(V;\vec{M}_{\vec{\omega}})+\log 2}{\log|\mathcal{V}|}.\tag{40}$$

Combining both with $I(V;\vec{M}_{\vec{\omega}})\leq\beta$ proves the lemma.

Proof of Lemma 6.

Let $N(\mathbb{G}_k^n,\rho,\delta)$ be the covering number of $\mathbb{G}_k^n$ with respect to metric ρ and covering radius δ. It is defined as

$$\min\{|\mathcal{V}|:\mathbb{G}_k^n\text{ admits a }\delta\text{-cover indexed by }\mathcal{V}\},\tag{41}$$

where a *δ-cover* is a set of points in $\mathbb{G}_k^n$ such that the union of their δ-balls contains $\mathbb{G}_k^n$. It is stated [19, Proposition 8] there are universal constants $s_1,s_2>0$ such that

$$(s_1\cdot D_g/\delta)^d\leq N(\mathbb{G}_k^n,\rho,\delta)\leq(s_2\cdot D_g/\delta)^d.\tag{42}$$

In addition, it is well-known that (e.g. [25, Equation 1.5]).

$$N(\mathbb{G}_k^n,\rho,\delta)\leq M(\mathbb{G}_k^n,\rho,\delta)\leq N(\mathbb{G}_k^n,\rho,\delta/2).\tag{43}$$

Putting two together we have

$$(s_1\cdot D_g/\delta)^d\leq M(\mathbb{G}_k^n,\rho,\delta)\leq(2s_2\cdot D_g/\delta)^d.\tag{44}$$

Setting $c_1=s_1$ and $c_2=2s_2$ proves the lemma.

Proof of Lemma 8.

Recall the generative model in section 2, where $\mathcal{P}=\{\mathbb{P}\}$ is a set of probabilities defined on $\mathbb{M}_k^{n,\vec{p}}$. For clarity we first derive the case when $\vec{M}$ is complete, and its generalization for $\vec{M}_{\vec{\omega}}$ naturally follows. Remark the following result.

Remark 10. *For any $\bar{M},\bar{M}'\in\mathbb{M}_k^{n,\vec{p}}$,*

$$D_{k\ell}(\mathbb{P}_{\bar{M}}(M)||\mathbb{P}_{\bar{M}'}(M))=\frac{1}{2\sigma^4}||\bar{M}-\bar{M}'||^2.\tag{45}$$

Proof. By the definition of KL divergence,

$$D_{k\ell}(\mathbb{P}_{\bar{M}}||\mathbb{P}_{\bar{M}'})=\mathbb{E}_{\mathbb{P}_{\bar{M}}}\log\left(\mathbb{P}_{\bar{M}}/\mathbb{P}_{\bar{M}'}\right),\tag{46}$$

where expectation $\mathbb{E}_{\mathbb{P}_{\bar{M}}}$ is taken over $\mathbb{P}_{\bar{M}}$. Further, by the definition of matrix-variate normal distribution (e.g. [6]),

$$\mathbb{P}_{\bar{M}}(\vec{M})=\exp\left(-||(\vec{M}-\bar{M}||^2/2\sigma^4\right)/\sqrt{(2\pi)^{np}}.\tag{47}$$

This implies

$$\log\frac{\mathbb{P}_{\bar{M}}}{\mathbb{P}_{\bar{M}'}}=-\frac{1}{2\sigma^4}(||\vec{M}-\bar{M}||^2-||\vec{M}-\bar{M}'||^2).\tag{48}$$

In addition,

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}_M}\,||\vec{M}-\bar{M}||^2-||\vec{M}-\bar{M}'||^2\\&=\mathbb{E}\sum_{i,j}(\vec{M}_{ij}-\bar{M}_{ij})^2-(\vec{M}_{ij}-\bar{M}'_{ij})^2\\&=\mathbb{E}\sum_{i,j}2\vec{M}_{ij}(\bar{M}'_{ij}-\bar{M}_{ij})+\bar{M}_{ij}^2-(\bar{M}'_{ij})^2\\&=\sum_{i,j}2\bar{M}_{ij}(\bar{M}'_{ij}-\bar{M}_{ij})+\bar{M}_{ij}^2-(\bar{M}'_{ij})^2\\&=\sum_{i,j}(\bar{M}'_{ij}-\bar{M}_{ij})^2=||\bar{M}'-\bar{M}||^2.\end{aligned}\tag{49}$$

Putting all together completes the proof. □

The arguments for Lemma 8 is almost the same as those for Remark 10, except now we have

$$\mathbb{P}_{\bar{M}|\vec{\omega}}(\vec{M}) = \prod\nolimits_{(i,j)\in\vec{\omega}} \tilde{\mathbb{P}}_{\bar{M}}(\vec{M}_{ij}), \tag{50}$$

which admits a matrix form (can be easily verified)

$$\mathbb{P}_{\bar{M}|\vec{\omega}}(\vec{M}) = \frac{\exp\left(-\frac{1}{2\sigma^4}||W_{\vec{\omega}} \circ (\vec{M} - \bar{M})||^2\right)}{\sqrt{(2\pi)^{|\vec{\omega}|}}}. \tag{51}$$

Keeping $W_{\vec{\omega}}$ through the derivation proves the lemma.

Proof of Theorem 9.

The first step is to apply the extended generalized Fano method (Lemma 5) to derive a lower bound. To do so, we need to fulfill its two conditions (22) and (23) respectively.

Let $\{\vec{M}_v \in \mathbb{M}_k^{n,\vec{p}}\}_{v\in\mathcal{V}}$ be a set inducing a 2δ-packing on $\mathbb{G}_k^n$, i.e. $\rho(\theta(\vec{M}_v), \theta(\vec{M}_{v'})) \geq 2\delta$ for any $v \neq v'$. This set fulfills (22). Note we have chosen all matrices from $\mathbb{M}_k^{n,\vec{p}}$, which does not weaken the analysis since this set can induce the entire $\mathbb{G}_k^n$ through θ.

It takes more effort to fulfill (23). Recall $I(V; \vec{M}_{\vec{\omega}})$ is based on the condition that input matrices share the same factor. Then our problem is equivalent to testing V using a single matrix $\vec{M}$ randomly drawn from $\mathbb{M}_k^{n,\vec{p}}$. This allows us to apply the generative model on $\mathbb{M}_k^{n,\vec{p}}$ in section 2.

For any $\vec{M} \in \mathbb{M}_k^{n,\vec{p}}$, let $\vec{A} \in \mathbb{R}^{k\times\vec{p}}$ be a loading in its factorization and $\vec{W} \in \mathbb{R}^{n\times\vec{p}}$ be its mask such that $\vec{W}_{ij} = 1$ if $\vec{M}_{ij}$ is observed and $\vec{W}_{ij} = 0$ otherwise. We have

$$\begin{aligned}
&I(V; \vec{M}_{\vec{\omega}}) \\
&\leq I(V; \vec{M}_{\vec{\omega}}|\vec{A}) \\
&\leq \mathbb{E}_{\vec{A}} \frac{1}{|\mathcal{V}|^2} \sum_{v,v'\in\mathcal{V}} D_{k\ell}(\mathbb{P}_v(\vec{M}|\vec{A})||\mathbb{P}_{v'}(\vec{M}|\vec{A})) \\
&\leq \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} \mathbb{E} D_{k\ell}(\mathbb{P}_v(\vec{M}|\vec{A})||\mathbb{P}_{v'}(\vec{M}|\vec{A})) \\
&= \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} \mathbb{E} \frac{1}{2\sigma^4} ||\vec{W} \circ (D_{\mathcal{G}_v} - D_{\mathcal{G}_{v'}})\vec{A}||^2 \qquad (52) \\
&\leq \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} \frac{||\vec{W}||^2}{2\sigma^4} ||D_{\mathcal{G}_v} - D_{\mathcal{G}_{v'}}||^2 \cdot \mathbb{E}||\vec{A}||^2 \\
&\leq \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} \frac{|\vec{\omega}|}{2\sigma^4} \cdot \rho(\mathcal{G}_v, \mathcal{G}_{v'}) \cdot \Sigma_{\vec{A}} \\
&\leq \frac{1}{2\sigma^4} |\vec{\omega}| \cdot \tau(\mathbb{G}_k^n) \cdot \Sigma_{\vec{A}},
\end{aligned}$$

where the first inequality is based on Lemma 7 where we condition both probabilities on a loading A; the second inequality is based on (29); the third inequality is due to the convexity of $\mathbb{E}_{\vec{A}}$; the first equality is based on Lemma 8; the fourth inequality is based on simple algebra argument (see Remark 14), and the fifth inequality is based on an extended argument of [20, Lemma A.1.2.] (see Remark 15); the last inequality is by the fact that $\rho(\mathcal{G}_v, \mathcal{G}_{v'}) \leq \tau(\mathbb{G}_k^n)$.

Till now we have fulfilled both conditions in the generalized Fano method described in Lemma 5. Together with Lemma 6 that bounds $|\mathcal{V}|$, this implies a lower bound

$$\frac{\delta}{2}\left(1 - \frac{|\vec{\omega}|\tau(\mathbb{G}_k^n)\Sigma_A/2\sigma^4 + \log 2}{d\left(\tau(\mathbb{G}_k^n)c/\delta\right)^d \log\left(\tau(\mathbb{G}_k^n)c/\delta\right)}\right). \tag{53}$$

As a standard strategy, it remains to choose a proper δ so that the 'big' fraction in (53) is upper bounded by $1/2$.

We are mainly interested in the order of parameters. First relax the lower order term $\log\left(\tau(\mathbb{G}_k^n)c/\delta\right) \geq \log(2c)$ since $\tau(\mathbb{G}_k^n) \geq 2\delta$ and the constant $\log 2 \geq \log 1 = 0$. Rearranging terms, we wish to choose a δ satisfying

$$\delta \geq \left(\frac{c^d \cdot d \cdot \log 2c \cdot \tau(\mathbb{G}_k^n)^d \cdot 2\sigma^4}{2|\vec{\omega}|\tau(\mathbb{G}_k^n)\Sigma_A}\right)^{1/d}. \tag{54}$$

Further, since $d = k(n-k)$ is a positive integer, it is easy to verify $d^{1/d} \leq 1.45$ and $(\log(2c))^{1/d} \leq \log(2c)$. Plugging both in the above lower bound and merging constants and terms depending on c into c', we have

$$\delta \geq c' \cdot \tau(\mathbb{G}_k^n)^{1-1/d} \cdot \left(|\vec{\omega}|\Sigma_A/\sigma^4\right)^{-1/d}. \tag{55}$$

Plugging this back to the lower bound (53) and merging constants again proves the theorem.

5.1 REMARKS

Remark 11. *Any $M, M' \in \mathbb{M}_k^{n\cdot}$ admit a joint full-rank factorization $M = QA$ and $M' = QA'$ for some $Q \in \mathbb{R}^{n\times k}$ if and only if they admit a joint factorization $M = DB$ and $M' = DB'$ for some $D \in \mathbb{S}_k^n$.*

Proof. A well known fact is that that every subspace of $\mathbb{R}^n$ admits an orthonormal basis (by Gram-Schmidt process). Thus for one direction, any Q induces a subspace $span(Q)$, which admits an orthonormal basis $D \in \mathbb{S}_k^n$. This means $Q = DL$ for some expressive coefficient matrix L and thus $M = QA = D(LA)$. The other direction is trivial. □

Remark 12. *Every $M \in \mathbb{M}_k^{n\cdot}$ admits exactly one $\mathcal{S} \in \mathbb{G}_k^n$ such that $\mathcal{S} = [D]$ for any $D \in \mathbb{S}_k^n$ satisfying $M = DA$.*

Proof. Given any two factorizations $M = DA = D'A'$ with $D, D' \in \mathbb{S}_k^n$, to justify the remark it suffices to show $span(D) \subseteq span(D')$ and $span(D) \supseteq span(D')$.

For the first direction, it suffices to find a $W \in \mathbb{R}^{k\times k}$ such that $D'W = D$. This is easy as $(D')^T D'$ is invertible since $D' \in \mathbb{S}_k^n$. (In fact, we only need D' to have full column rank.) Thus one can set $W = ((D')^T D')^{-1}(D')^T D$. Similar arguments apply for the other direction. □

Remark 13. *In Proposition 1, if one always has* $\theta(M) = \theta(M')$*, then* $\mathfrak{M}(\hat{\theta}) \geq \delta \cdot \Pr\{\hat{V}(\vec{M}_{\vec{\omega}}) \neq V\}$.

Proof. Condition $\theta(M) = \theta(M')$ implies we can fix $V = V'$ while designing the collective hypothesis testing problem. This means $\Pr\{V \neq V'\} = 0$ and $\Pr\{V = V'\} = 1$. Plugging both into the proof of Proposition 1 (last inequality) justifies the remark. □

Remark 14. *For any same sized matrices* A *and* B*, we have* $||A \circ B||^2 \leq ||A||^2 \cdot ||B||^2$.

Proof. For all sums taken over all matrix indices, it follows $||A \circ B||^2 = \sum (A_{ij}B_{ij})^2 = \sum (A_{ij})^2 (B_{ij})^2 \leq \sum (A_{ij})^2 \sum (B_{i'j'})^2 = ||A||^2 ||B||^2$, where the inequality is by the fact that $(B_{ij})^2 \leq \sum (B_{i'j'})^2$ for any (i, j). □

Remark 15. *For any* $D, D' \in \mathbb{S}_k^n$,

$$||D - D'||^2 \leq ||DD^T - D'(D')^T||^2. \tag{56}$$

Proof. This remark is a matrix extension of [20, Lemma A.1.2]. Let $D_{:j}$ denote the column j of D. We have

$$\begin{aligned} &||DD^T - D'(D')^T||^2 \\ &= \sum_j ||D_{:j}D_{:j}^T - D'_{:j}(D'_{:j})^T||^2 \\ &\leq \sum_j ||D_{:j} - D'_{:j}||^2 = ||D - D'||^2, \end{aligned} \tag{57}$$

where the inequality is based on [20, Lemma A.1.2] and the fact that $||D_{:j} - D'_{:j}|| \leq \sqrt{2}$ since $D, D' \in \mathbb{S}_k^n$. □

6 SIMULATION

In this section we empirically evaluate the learning rate of CMF under two settings, one without negative transfer and the other with negative transfer (NT):

NT-Free: in this case, we randomly generate a factor $D \in \mathbb{S}_k^n$ and two loadings $A \in \mathbb{R}^{n \times p}$ and $A' \in \mathbb{R}^{n \times p'}$ to construct matrices $M = DA$ and $M' = D'A'$. By this means, M, M' are guaranteed to share the same factor and negative transfer does not exist.

NT-Likely: in this case, we randomly and independently generate two factors $D, D' \in \mathbb{S}_k^n$ and two loadings A, A' same sized as in the NT-Free case. The two matrices are constructed by $M = DA$ and $M' = D'A'$. By this means, it is likely $[D] \neq [D']$ and thus negative transfer exists.

In evaluation we simply set n, p, p' to 50 and set k to 10. To examine the learning rate, the ratio of observations, denoted by r, is varied from 0.1, 0.3, 0.5, 0.7 to 0.9. At each choice of r, we randomly select $r \cdot np$ number of entries in each matrix to form the observation $\vec{M}_{\vec{\omega}}$. The CMF algorithm in [17] is implemented with no use of prediction link function. After performing CMF on $\vec{M}$, the loss function in the proposition is evaluated. Rewrite its notation as

$$Loss_{\omega}(\hat{\theta}) = \frac{1}{2}\left[\rho(\hat{\theta}, \theta(M)) + \rho(\hat{\theta}, \theta(M'))\right]. \tag{58}$$

In addition, fpr each r we repeat the random choice of $\vec{M}_{\vec{\omega}}$ for 10 times and report the averaged loss in Figure 3.

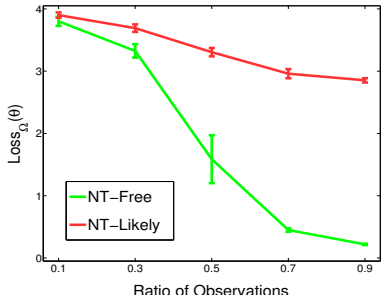

Figure 3: Performance simulation.

From Figure 3 it is clear that CMF converges much slower when negative transfer exists, as compared with the case when negative transfer does not exist. The bias is also quite obvious. These coincide with our theoretical discoveries.

7 CONCLUSION AND DISCUSSION

This paper presents a first theoretical explanation of negative transfer in collective matrix factorization. We present a min-max lower bound of the CMF estimator and show negative transfer gives rise to an additional bias term that depends only on the structure of the factor space. We further present a finer lower bound and show negative transfer slows the learning rate from $\Omega(|\vec{\omega}|^{-1})$ to $\Omega(|\vec{\omega}|^{-1/d})$, where d is the dimension of Grassmannian containing the subspaces spanned by matrix factors.

A limitation of this study is we assumed full-rank factorization. As suggested by the theory, increasing k may mitigate negative transfer, but clearly at the cost of increasing estimation variance. How these two aspects trade with each other remains unclear, even though our analysis may be naively extended for a larger k' (by simply basing everything on $\mathbb{G}_{k'}^n$). In addition, in reality two matrices may have different ranks and their induced subspaces may partly overlap [10]. This partial overlapping is merely implicitly captured in our analysis (thorough the choice of δ) and stronger results may be obtained by explicitly modeling it.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. This work was supported in part by the National Natural Science Foundation of China under Grant No.61232001 and No.61420106009 for Jianxin Wang and NSF grant 1513324 to Jun Huan.

References

[1] Deepak Agarwal, Bee-Chung Chen, and Bo Long. Localized factor models for multi-context recommendation. In *SIGKDD*, 2011.

[2] Guillaume Bouchard, Dawei Yin, and Shengbo Guo. Convex collective matrix factorization. In *AISTATS*, 2013.

[3] Thomas Franz, Antje Schultz, Sergej Sizov, and Steffen Staab. Triplerank: Ranking semantic web data by tensor decomposition. In *International Semantic Web Conference (ISWC)*, 2009.

[4] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[5] Suriya Gunasekar, Makoto Yamada, Dawei Yin, and Yi Chang. Consistent collective matrix completion under joint low rank structure. In *AISTATS*, 2015.

[6] Arjun K Gupta and Daya K Nagar. *Matrix variate distributions*, volume 104. CRC Press, 1999.

[7] David Cohn Thomas Hofmann. The missing link-a probabilistic model of document content and hypertext connectivity. In *NIPS*, 2001.

[8] Alexander Jung, Yonina Eldar, and Norbert Gortz. On the minimax risk of dictionary learning. *Information Theory, IEEE Transactions on*, 62(3):1501–1515, 2016.

[9] Raghunandan H Keshavan, Sewoong Oh, and Andrea Montanari. Matrix completion from a few entries. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pages 324–328. IEEE, 2009.

[10] Arto Klami, Guillaume Bouchard, Abhishek Tripathi, et al. Group-sparse embeddings in collective matrix factorization. In *Proceedings of International Conference on Learning Representations (ICLR) 2014*, 2014.

[11] Christoph Lippert, Stefan Hagen Weber, Yi Huang, Volker Tresp, Matthias Schubert, and Hans-Peter Kriegel. Relation prediction in multi-relational domains using matrix factorization. In *NIPS Workshop: Structured Input-Structured Output*, 2008.

[12] Bo Long, Zhongfei Mark Zhang, Xiaoyun Wu, and Philip S Yu. Spectral clustering for multi-type relational data. In *ICML*, 2006.

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

[14] Amritanshu Prasad. Counting subspaces of a finite vector space. *Resonance*, 15(11):977–987, 2010.

[15] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887. ACM, 2008.

[16] Ajit Singh and Geoffrey Gordon. A Bayesian matrix factorization model for relational data. In *UAI*, 2010.

[17] Ajit P Singh and Geoffrey J Gordon. Relational learning via collective matrix factorization. In *SIGKDD*, 2008.

[18] Ji-Guang Sun. Perturbation of angles between linear subspaces. *Journal Of Computational Mathematics*, 5(1), 1987.

[19] Stanisław Szarek. Nets of Grassmann manifold and orthogonal group. In *Proceedings of Banach Space Workshop, University of Iowa Press*, pages 169–185, 1982.

[20] Vincent Q Vu and Jing Lei. Minimax rates of estimation for sparse PCA in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 1278–1286, 2012.

[21] Kenan Y Yılmaz, Ali T Cemgil, and Umut Simsekli. Generalised coupled tensor factorisation. In *NIPS*, 2011.

[22] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

[23] Kai Yu, Shipeng Yu, and Volker Tresp. Multi-label informed latent semantic indexing. In *SIGIR*, 2005.

[24] Yu Zhang, Bin Cao, and Dit-Yan Yeung. Multi-domain collaborative filtering. In *UAI*, 2010.

[25] Ding-Xuan Zhou. Capacity of reproducing kernel spaces in learning theory. *Information Theory, IEEE Transactions on*, 49(7):1743–1752, 2003.

[26] Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye. From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records. In *SIGKDD*, 2014.