BINARY MATRIX COMPLETION WITH PERFORMANCE GUARANTEES FOR SINGLE INDIVIDUAL HAPLOTYPING

Somsubhra Barik and Haris Vikalo

Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, Texas 78712

ABSTRACT

We study the problem of approximating a partially observed matrix by a product of two low-rank matrices where the data as well as the factors are constrained to be binary. This computationally challenging task is motivated by the single individual haplotyping problem which attracted considerable attention in computational biology and is of critical importance for personalized medicine applications. We analyze a binary-constrained variant of the alternating minimization algorithm for solving the aforementioned problem in the scenario where the matrices are rank-one, establish its performance and convergence properties, and in doing so provide the first theoretical guarantees for haplotype reconstruction expressed in terms of the minimum error-correction score. Sample complexity required for reconstruction is derived and experiments are performed on both synthetic and real datasets, demonstrating superiority of the proposed framework over competing methods.

Index Terms— matrix completion, single individual haplotyping, sparsity, alternating minimization

1. INTRODUCTION

Finding a rank-k approximation $\mathbf{M} \in \mathbb{R}^{m\times n}$, $k < \min\{m, n\}$, to a partially observed matrix is often reduced to the search for factors $\mathbf{U} \in \mathbb{R}^{m\times k}$ and $\mathbf{V} \in \mathbb{R}^{n\times k}$ such that $\mathbf{M} = \mathbf{U}\mathbf{V}^T$ [1, 2, 3, 4, 5]. In a host of applications, factors may exhibit structural properties such as sparsity, non-negativity or discreteness. Such applications include blind source separation [6], gene network inference [7], and clustering with overlapping clusters [8], to name a few. In this paper, we consider the rank-one decomposition of a binary matrix $\mathbf{M} \in \{0, 1\}^{m\times n}$ from its partial observations that are perturbed by bit-flipping noise. This problem belongs to a broader category of non-negative matrix factorization [2] or, more specifically, binary matrix factorization [9, 10, 11, 12]. Related prior work includes [9, 10] which considers decomposition of a binary $\mathbf{M}$ in terms of non-binary $\mathbf{U}$ and $\mathbf{V}$, while [11] explores a Bayesian approach to factorizing matrices having binary components. The approach in [12] constrains $\mathbf{M}$, $\mathbf{U}$ and $\mathbf{V}$ to all be binary; however, it requires a fully observed input matrix $\mathbf{M}$. Our framework, motivated by the *single individual haplotyping* (SIH) problem in computational biology, employs alternating minimization to compute factors from the partial noisy observations of $\mathbf{M}$ while imposing constraints to ensure specific structure of the factors.

The motivating application, single individual haplotyping from high-throughput DNA sequencing data, is an NP-hard problem concerned with reconstruction of the variations between chromosomes in an organism. We focus on diploid organisms (e.g., humans) whose DNA is organized in pairs of chromosomes. The chromosomes in a pair encode the same genetic information and are almost identical but differ from each other in a fraction of positions due to point mutations referred to as single nucleotide polymorphisms (SNPs). The sequence of SNPs on each chromosome in a pair is referred to as haplotype. Haplotype information is of critical importance for personalized medicine applications including the discovery of an individual's susceptibility to diseases [13], whole genome association studies [14], gene detection under positive selection, and the discovery of recombination patterns [15]. High-throughput DNA sequencing enables inference of haplotypes by providing information about short subsequences of the corresponding chromosomes; one can think of the sequencing data as being obtained by randomly sampling (with replacement) short substrings from each chromosome. Origin of sampled substrings (so-called *reads*) is not known a priori, i.e., it is unknown from which of the two chromosomes in a pair any given read was sampled. Single individual haplotyping essentially needs to partition the reads into two clusters, one for each chromosome in a pair, and use them to reconstruct the corresponding haplotypes. Low frequency of SNPs in humans (1 in 300 bases [16]), relatively short reads and the presence of sequencing error ($10^{-3}-10^{-2}$ error rates) render the SIH problem computationally challenging.

A widely used metric for characterizing the quality of haplotype assembly is the minimum error correction (MEC) score – essentially, the most likely number of sequencing errors (see Section 3 for a formal definition). Most prior work on SIH is focused on optimizing the MEC score [17]. These include the branch-and-bound approach in [18], greedy ap-

proach in [19], max-cut based formulation [20], MCMC [21], greedy-cut based [22] and flow-graph based approaches [22].

In this paper, we formulate SIH as a rank-one matrix completion problem and propose a binary-constrained variant of alternating minimization to solve it. We analyze the performance and convergence properties of the proposed algorithm, and provide the first theoretical guarantees for haplotype reconstruction expressed in a form of the bound on MEC score. Furthermore, we determine the sample complexity (essentially, the sequencing coverage) sufficient for the algorithm to converge. Experiments performed on both synthetic and real datasets demonstrate superiority of the proposed framework over competing methods. Matrix factorization framework was previously used to solve the SIH problem via gradient descent in [23] but the method there does not provide performance guarantees established in the current paper.

2. SYSTEM MODEL AND PROBLEM STATEMENT

To set up mathematical framework for our problem, we first provide the following definition of incoherence (due to [3]).

Definition 1. *A rank-k matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ with singular value decomposition $\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^T$ is incoherent with parameter μ if for every $1 \le i \le m$, $1 \le j \le n$,*

$$\|\mathbb{P}_{\mathbf{U}}(\mathbf{e}_i)\|_2 \le \frac{\mu\sqrt{k}}{\sqrt{m}}, \text{ and } \|\mathbb{P}_{\mathbf{V}}(\mathbf{e}_j)\|_2 \le \frac{\mu\sqrt{k}}{\sqrt{n}}.$$

Here $\mathbb{P}_{\mathbf{U}}(\cdot)$ denotes the projection operator onto the column subspace of $\mathbf{U}$, and $\mathbf{e}_i$ is the i^{th} standard basis vector.

Let n denote the number of sequencing reads used in reconstruction of a haplotype of length m. We organize the reads into an $m \times n$ SNP fragment matrix $\mathbf{R}$ whose j^{th} column R_j contains information provided by the j^{th} read. Since diploid organisms typically have bi-allelic chromosomes (i.e., only two out of four nucleotides are possible at each SNP position), ± 1 labels can be ascribed to the entries of $\mathbf{R}$ that provide SNP information, where the mapping between nucleotides and binary labels follows arbitrary convention. Since reads are typically much shorter than haplotypes, many entries of $\mathbf{R}$ are uninformative. Let Ω denote the set of informative entries of $\mathbf{R}$, i.e., the set of (i, j) such that the j^{th} read covers the i^{th} SNP. Define an operator $P_\Omega(\cdot)$ as

$$[P_\Omega(\mathbf{R})]_{ij} = \begin{cases} R_{ij}, & (i,j) \in \Omega \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Therefore, $P_\Omega(\mathbf{R})$ is a matrix with entries in $\{-1, 0, 1\}$. Let $\mathcal{H} = \{h_1, h_{-1}\}$ denote the pair of haplotype sequences of a diploid organism, with $h_i \in \{-1, 1\}^m, i = \pm 1$. Note that $h_1 = -h_{-1}$. $P_\Omega(\mathbf{R})$ can be thought of as being obtained by sampling, with errors, a rank-one matrix $\mathbf{M}$ having ± 1 entries. Moreover, $\mathbf{M} = \hat{\mathbf{u}}^\star(\hat{\mathbf{v}}^\star)^T = \sigma^\star\mathbf{u}^\star(\mathbf{v}^\star)^T$ where $\hat{\mathbf{u}}^\star$ and $\hat{\mathbf{v}}^\star$ are vectors with ± 1 entries and have lengths m and n, respectively, $\mathbf{u}^\star$ and $\mathbf{v}^\star$ are normalized $\hat{\mathbf{u}}^\star$ and $\hat{\mathbf{v}}^\star$, and $\sigma^\star > 0$ is the singular value of $\mathbf{M}$. Note that $\hat{\mathbf{u}}^\star$ represents the haplotype h_1 or h_{-1} (the assignment is arbitrary) and $\hat{v}_j^\star$ indicates the membership of the j^{th} read, i.e., $\hat{v}_j^\star = i$ implies that the j^{th} read is sampled from h_i. Hence the SIH problem can be formalized as the optimization

$$(\hat{\mathbf{u}}, \hat{\mathbf{v}}) = \underset{\mathbf{u}\in\{1,-1\}^m, \mathbf{v}\in\{1,-1\}^n}{\arg\min} f(\mathbf{u}, \mathbf{v}), \tag{2}$$

where the loss function $f(\mathbf{u}, \mathbf{v})$ is often chosen to be

$$f(\mathbf{u}, \mathbf{v}) = \|P_\Omega(\mathbf{R} - \mathbf{u}\mathbf{v}^T)\|_F^2 = \sum_{(i,j)\in\Omega} (R_{ij} - u_i v_j)^2.$$

For the analysis in Section 3, we need to impose certain assumptions on the bit-flipping noise matrix $\mathbf{N}$. Let p_e denote the sequencing error probability. Then

$$N_{ij} = \begin{cases} 0, & \text{with probability } (1 - p_e), \\ -2M_{ij}, & \text{with probability } p_e. \end{cases} \tag{3}$$

We assume $\mathbf{N}$ is full rank with the SVD $\mathbf{N} = \mathbf{U}^N\Sigma^N(\mathbf{V}^N)^T$, $\mathbf{U}^N \in \mathbb{R}^{m\times m}$, $\mathbf{V}^N \in \mathbb{R}^{n\times m}$, $\Sigma^N = \text{diag}(\sigma_1^N, \ldots, \sigma_m^N)$, and $m \le n$. The observed data (the SNP fragment matrix) can then be modeled as $\mathbf{R} = \mathbf{M} + \mathbf{N}$. Note that $\mathbf{N}$ fits the worst-case noise model of [24, 25]. There, the entries of $\mathbf{N}$ are assumed to be distributed arbitrarily with the restriction that there exists an entry-wise uniform upper bound on the absolute value, i.e., $|N_{ij}| \le N_{\max}$, leading to $\|N\|_F \le \sqrt{mn}N_{\max}$. However, in our problem the noise matrix has additional properties – namely, the entries of $\mathbf{N}$ are Bernoulli variables with parameter p_e. The following lemma provides a bound on the spectral norm of the partially observed noise matrix $P_\Omega(\mathbf{N})$ (the proof is omitted for brevity).

Lemma 1. *Let $\mathbf{N}$ be an $m \times n$ sequencing error matrix defined in (3). Let Ω be the sample set of the observed entries and let p be the observation probability. If p_e denotes the sequencing error rate, then with high probability it holds that*

$$\|P_\Omega(\mathbf{N})\|_2/p \le 2N_{\max}p_e\sqrt{mn}.$$

3. THE ALGORITHM AND ITS ANALYSIS

A straightforward application of alternating minimization to solving (2) involves relaxing binary constraints so that $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{v} \in \mathbb{R}^n$, and performing updates to one vector while keeping the other one fixed,

$$\begin{aligned} &\hat{\mathbf{v}} \leftarrow \arg\min_{\mathbf{v}\in\mathbb{R}^n} \sum_{(i,j)\in\Omega} (R_{ij} - \hat{u}_i v_j)^2, \text{ and} \\ &\hat{\mathbf{u}} \leftarrow \arg\min_{\mathbf{u}\in\mathbb{R}^m} \sum_{(i,j)\in\Omega} (R_{ij} - u_i \hat{v}_j)^2, \text{ and repeat.} \end{aligned} \tag{4}$$

Once a termination condition is met, the entries of $\hat{\mathbf{u}}$ would need to be rounded to ± 1 to estimate the haplotype vector $\mathbf{u}^\star$. Note that initialization heavily impacts the performance of alternating minimization. The singular vector corresponding to the largest singular value of $P_\Omega(\mathbf{R})$ is a suitable choice. To avoid computationally expensive singular value decomposition, one can rely on the efficient power iteration method to compute this singular vector [23].

To guarantee convergence of the alternating minimization in (4), $\hat{\mathbf{u}}^t$ and $\hat{\mathbf{v}}^t$ need to be incoherent (see Definition 1) in each iteration t [5]. To ensure this in the initial step, one may "clip" (set to zero) entries of $\hat{\mathbf{u}}^0$ that exceed a certain threshold. The singular vector obtained by power iterations minimizes the distance from $\mathbf{u}^\star$; the clipping makes sure that the information is spread across all the dimensions of $\hat{\mathbf{u}}$ as opposed to being concentrated in only few entries.

The updates (4) ignore the fact that the true factors $\mathbf{u}$ and $\mathbf{v}$ consist of discrete ± 1 entries; instead, the previously described procedure imposes binary constraints only after completing the iterations. This may adversely impact the convergence of alternating minimization; to see this, note that when e.g. $\hat{\mathbf{v}}$ is updated according to (4), its j^{th} entry is found as

$$\hat{v}_j^{(t+1)} = \arg\min_{v\in\mathbb{R}} \sum_{i|(i,j)\in\Omega} (R_{ij} - \hat{u}_i^{(t)} v)^2 = \frac{\sum\limits_{i|(i,j)\in\Omega} R_{ij}\hat{u}_i^{(t)}}{\sum\limits_{i|(i,j)\in\Omega} \left(\hat{u}_i^{(t)}\right)^2}.$$

We empirically observe that as the iterations progress $\hat{v}_j^{(t+1)}$ may become very large or very small, which leads to potential loss of incoherence of the iterates. To maintain incoherence, it is desirable that the entries of $\hat{\mathbf{u}}^{(t)}$ and $\hat{\mathbf{v}}^{(t)}$ remain close to ± 1. To this end, we impose the inherent binary structure of $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ to (4), arriving at updates

$$\hat{v}_j^{(t+1)} = \begin{cases} 1 & \text{if } \sum\limits_{i|(i,j)\in\Omega} R_{ij}\hat{u}_i^{(t)} \geq 0, \\ -1 & \text{otherwise,} \end{cases} \tag{5}$$

and a similar expression for $\hat{u}_j^{(t+1)}$. In other words, we project the solution of each step onto $\{1,-1\}^m$, i.e., replace the continuous minimization step with a discrete counterpart

$$\begin{aligned} \hat{\mathbf{v}} &\leftarrow \arg\min_{\mathbf{v}\in\{1,-1\}^n} \sum_{(i,j)\in\Omega} (R_{ij} - \hat{u}_i v_j)^2, \text{ and} \\ \hat{\mathbf{u}} &\leftarrow \arg\min_{\mathbf{u}\in\{1,-1\}^m} \sum_{(i,j)\in\Omega} (R_{ij} - u_i \hat{v}_j)^2. \end{aligned} \tag{6}$$

The binary constrained alternating minimization algorithm is formalized as Algorithm 1.

The non-differentiability of (5), however, makes the analysis of convergence of Algorithm 1 intractable. To remedy this, we approximate (5) using a logistic function $f(x) = (e^x - 1)/(e^x + 1)$, thus replacing the $\hat{\mathbf{v}}$ and $\hat{\mathbf{u}}$ updates in Algorithm 1 by

$$\begin{aligned} \hat{v}_j^{(t+1)} &= \frac{\exp\left(\frac{1}{m}\sum_{i|(i,j)\in\Omega} R_{ij}u_i^{(t)}\right) - 1}{\exp\left(\frac{1}{m}\sum_{i|(i,j)\in\Omega} R_{ij}u_i^{(t)}\right) + 1}, \text{ and} \\ \hat{u}_i^{(t+1)} &= \frac{\exp\left(\frac{1}{n}\sum_{j|(i,j)\in\Omega} R_{ij}v_j^{(t+1)}\right) - 1}{\exp\left(\frac{1}{n}\sum_{j|(i,j)\in\Omega} R_{ij}v_j^{(t+1)}\right) + 1}, \end{aligned} \tag{7}$$

where $1 \leq j \leq n$, $1 \leq i \leq m$, and $\mathbf{u}^t$ and $\mathbf{v}^t$ denote normalized $\hat{\mathbf{u}}^t$ and $\hat{\mathbf{v}}^t$. The equations (7) relax the binary constraints on $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ while ensuring that the values remain bounded within the interval $[1,-1]$. Note that in our multiple tests on both synthetic and experimental data, approximation (7) did not lead to any noticeable loss of performance.

Algorithm 1 Binary-Constrained Alt-Min for SIH

Require: $P_\Omega(\mathbf{R}) \in \{0,1,-1\}^{m\times n}$, $\Omega \subseteq [m]\times[n]$, p

Power Iteration: Generate $\mathbf{u}^0$ (top singular vector of $P_\Omega(\mathbf{R})/p$)

Clipping: Set entries of $\mathbf{u}^0$ greater than $\frac{2}{\sqrt{m}}$ to zero

for $t = 0, 1, 2, \ldots, T-1$ **do**

$$\hat{\mathbf{v}}^{(t+1)} \leftarrow \arg\min_{\mathbf{v}\in\{1,-1\}^n} \sum_{(i,j)\in\Omega} \left(R_{ij} - \hat{u}_i^{(t)} v_j\right)^2$$

$$\hat{\mathbf{u}}^{(t+1)} \leftarrow \arg\min_{\mathbf{u}\in\{1,-1\}^m} \sum_{(i,j)\in\Omega} \left(R_{ij} - u_i \hat{v}_j^{(t+1)}\right)^2$$

end for

Output: $\hat{\mathbf{u}}^{(T)}$ is the estimate $\hat{\mathbf{u}}$ of the haplotype vector

Theorem 1. *Let $\hat{\mathbf{u}}^\star \in \{1,-1\}^m$ and $\hat{\mathbf{v}}^\star \in \{1,-1\}^n$ denote the haplotype and read membership vectors, respectively, and let $\mathbf{R} = \mathbf{M} + \mathbf{N}$ denote the observed SNP fragment matrix where $\mathbf{M} = \hat{\mathbf{u}}^\star(\hat{\mathbf{v}}^\star)^T = \mathbf{u}^\star\sigma^\star(\mathbf{v}^\star)^T$, $\mathbf{N}$ is the noise matrix with $N_{\max}$ and p_e as defined in (3), $\mathbf{u}^\star$ and $\mathbf{v}^\star$ are normalized versions of $\hat{\mathbf{u}}^\star$ and $\hat{\mathbf{v}}^\star$, respectively, and $\sigma^\star$ is the singular value of $\mathbf{M}$. Let $\alpha = n/m \geq 1$. Assume that each entry of $\mathbf{M}$ is observed uniformly randomly with probability*

$$p > C\frac{\sqrt{\alpha}}{m\delta_2^2}\log n \log\left(\frac{\|\mathbf{M}\|_F}{\epsilon}\right)\left(p_e + \frac{64}{3}\delta_2\right), \tag{8}$$

where $\delta_2 \in \left[0, \frac{1}{21}(3.93 - C'N_{\max}p_e)\right]$ and $C, C' > 0$ are global constants. Then, for any $\epsilon > 0$, after $T = \mathcal{O}(\log(\|\mathbf{M}\|_F/\epsilon))$ iterations of Algorithm 1, the estimate $\hat{\mathbf{M}}^{(T)}$ with high probability satisfies

$$\|\mathbf{M} - \hat{\mathbf{M}}^{(T)}\|_F \leq \epsilon + 16\frac{p_e\sigma^\star}{3\delta_2}(2 + (2 + 3N_{\max})\delta_2). \tag{9}$$

The proof of Theorem 1 relies on demonstrating a geometric decay of the distance between the subspace spanned

by $\hat{\mathbf{u}}^{(t)}$ and the one spanned by $\mathbf{u}^\star$ (similarly for $\hat{\mathbf{v}}^{(t)}$ and $\mathbf{v}^\star$), details are omitted for brevity. The following corollary follows directly from Theorem 1.

Corollary 1. *Under the conditions of Theorem 1, the normalized Minimum Error Correction score with respect to* $\mathbf{R}$*, defined as* $\tilde{MEC} = \frac{1}{mn}\|P_\Omega(\mathbf{R} - \hat{\mathbf{M}}^{(T)})\|_0$*, satisfies*

$$\tilde{MEC}(\hat{\mathbf{M}}^{(T)}) \leq \frac{\epsilon}{\sqrt{mn}} + \frac{16p_e}{3\delta_2}(2 + (2 + 3N_{\max})\delta_2) + \frac{1}{\sqrt{mn}}\|P_\Omega(\mathbf{N})\|_F. \quad (10)$$

Theorem 1 and Corollary 1 imply that for a given error probability p_e, if the sample probability p satisfies the condition (8), then Algorithm 1 can minimize the MEC score up to some noise factors in $\mathcal{O}(\log(\|\mathbf{M}\|_F/\epsilon))$ iterations. The corresponding sample complexity, i.e., the number of entries of $\mathbf{R}$ needed for the recovery of $\mathbf{M}$ is $|\Omega| = \mathcal{O}\left(\frac{\sqrt{\alpha}}{\delta_2^2} n \log n \log\left(\frac{\|\mathbf{M}\|_F}{\epsilon}\right)\left(p_e + \frac{64}{3}\delta_2\right)\right)$. Note that compared to (9), expression (10) has an additional noise term. This is due to the fact that unlike the loss function $\|\mathbf{M} - \hat{\mathbf{M}}^{(T)}\|_F$ in (9), the MEC score of $\hat{\mathbf{M}}^{(T)}$ is calculated with respect to the observed matrix $P_\Omega(\mathbf{R})$.

4. EXPERIMENTS

We test our algorithm on the experimental dataset containing Fosmid pool-based next generation sequencing data for HapMap trio child $NA12878$ [22] and compare its performance with the structurally-constrained gradient descent (SCGD) approach in [23] and another recent SIH software ProbHap [26] shown to be superior to several prior methods [20, 22, 27]. The Fosmid dataset is characterized by very long fragments, high SNP to read ratio, and sequencing coverage of about 3X. Table 1 shows the MEC rate (average number of mismatches per SNP position across the reads) and runtimes for 9 of the chromosomes. As seen there, our algorithm outperforms other methods for majority of the chromosomes shown; it is second best in terms of runtime (behind SCGD).

While the MEC score is essential for characterizing performance of haplotype assembly in practice, it is ultimately a proxy for the *reconstruction rate* [28]. Recall that the $\mathcal{H} = \{h_1, h_{-1}\}$ is the set of true haplotypes; let us denote the set of estimated haplotypes by $\hat{\mathcal{H}} = \{\hat{h}_1, \hat{h}_{-1}\}$. The reconstruction rate of $\hat{\mathcal{H}}$ with respect to $\mathcal{H}$ is defined as $\mathcal{R}_{\mathcal{H},\hat{\mathcal{H}}} = 1 - \frac{1}{2m}\min\left\{D(h_1, \hat{h}_1 + D(h_2, \hat{h}_2),\ D(h_1, \hat{h}_2) + D(h_2, \hat{h}_1)\right\}$, where $D(h_i, h_j) = \sum_{l=1}^{m} d(h_i(l), h_j(l))$ denotes the generalized Hamming distance between h_i and h_j, $h_i(l)$ is the l^{th} entry of $h_i, \forall\ l = 1, \ldots, m$, and the distance measure d is defined as $d(x, y) = 1$ if $x \neq 0, y \neq 0, x \neq y,\ 0$ otherwise, for any $x, y \in \{-1, 1, 0\}$.

We test the reconstruction rate performance of our method on the broadly used benchmarking dataset in [28] and compare[1] it with that of the previous work [23], HGHap [29] and MixSIH [27]. The results, obtained by averaging over 100 simulation runs for each combination of error rate and sequencing coverage, are reported in Table 2. As evident from the results, our method is either the best or the second best in all of the scenarios.

Table 1. MEC rates and runtimes on Fosmid dataset.

Chr	Algo. 1		SCGD		ProbHap	
	MEC	time(s)	MEC	time(s)	MEC	time(s)
1	0.034	65.0	0.04	44.2	0.058	87.7
2	0.035	71.6	0.035	49.5	0.055	88.9
3	0.034	61.1	0.036	41.5	0.057	84.3
4	0.029	60.7	0.034	41.8	0.053	67.1
5	0.032	52.9	0.036	39.9	0.054	64.6
20	0.044	18.1	0.044	13.0	0.055	30.9
21	0.035	11.5	0.041	8.5	0.051	15.6
22	0.054	11.7	0.055	8.6	0.061	31.4

Table 2. Reconstruction rate comparison on simulated data. Boldface values indicate best performance.

Error Rate	Cov.	Algo. 1	SCGD	HGHap	MixSIH
0.1	3X	**0.935**	0.869	0.934	0.775
0.1	5X	0.979	0.951	**0.990**	0.942
0.1	8X	**0.996**	**0.996**	0.987	0.972
0.1	10X	**0.999**	**0.999**	0.997	0.993
0.2	3X	**0.735**	0.677	0.677	0.68
0.2	5X	0.864	0.785	**0.91**	0.774
0.2	8X	**0.943**	0.899	0.884	0.932
0.2	10X	0.966	0.934	0.894	**0.969**

5. CONCLUSION

Motivated by the single individual haplotyping problem from computational biology, we proposed and analyzed a binary-constrained variant of the alternating-minimization algorithm for solving the rank one matrix factorization problem. We provided theoretical guarantees on the performance of the algorithm and analyzed its required sample probability; the latter has important implications on experimental specifications, namely, sequencing coverage. Performance of haplotype reconstruction is often expressed in terms of the minimum error correction score; we establish theoretical guarantees on the achievable MEC score for the proposed binary-constrained alternating minimization. Experiments with a real-world dataset as well as those with a widely used benchmarking simulated dataset demonstrated efficacy of our approach.

[1]The comparison with ProbHap is not shown since for synthetic data that algorithm returns haplotypes with a large fraction of SNPs missing.

6. REFERENCES

[1] Netflix, "Netflix prize," in *http://www.netflixprize.com/*. [Onine].

[2] H. Kim and H. Park, "Nnonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," in *SIAM J. Matrix Anal. Appl.*, 2008, vol. 30(2), pp. 713–730.

[3] E.J. Candès and B. Recht, "Exact matrix completion via convex optimization," in *Foundations of Computational mathematics*, 2009, vol. 9.6, pp. 717–772.

[4] B. Recht, M. Fazel, and P.A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," in *SIAM Review*, 2010, vol. 52(3), pp. 471–501.

[5] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *ArXiv e-prints*, 2012.

[6] A J. van der Veen, "Analytical method for blind binary signal separation," in *IEEE Signal Processing*, 1997, vol. 45, pp. 1078–1082.

[7] R. Liao, J. C.and Boscolo, Y. L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury, "Network component analysis: reconstruction of regulatory signals in biological systems," in *Proceedings of the National Academy of Sciences*, 2003, pp. 15522–15527.

[8] A. Banerjee, C. Krumpelman, J. Ghosh, S. Basu, and R. J. Mooney, "Model-based overlapping clustering," in *SIGKDD*. ACM, 2005, pp. 532–537.

[9] A. I. Schein, L. K. Saul, and L. H. Ungar, "A generalized linear model for principal component analysis of binary data.," in *AISTATS*, 2003, vol. 3(9), p. 10.

[10] A. Kabán and E. Bingham, "Factorisation and denoising of 0–1 data: a variational approach.," in *Neurocomputing*, 2008, vol. 71(10), pp. 2291–2308.

[11] E. Meeds, Z. Ghahramani, R. M. Neal, and S. T. Roweis, "Modeling dyadic data with binary latent factors.," in *NIPS*, 2006, vol. 3(9), pp. 977–984.

[12] Z. Zhang, T. Li, C. Ding, and X. Zhang, "Binary matrix factorization with applications.," in *ICDM*. IEEE, 2007, pp. 391–400.

[13] A. G. Clark, "The role of haplotypes in candidate gene studies," in *Genetic epidemiology*, 2004, vol. 27(4), pp. 321–333.

[14] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, and P. K. H. Tam, "The international hapmap project," in *Nature*, 2003, vol. 426(6968), pp. 789–796.

[15] P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner, et al., "Detecting recent positive selection in the human genome from haplotype structure," in *Nature*, 2002, vol. 419(6909), pp. 832–837.

[16] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, et al., "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," in *Nature*, 2001, vol. 409(6822), pp. 928–933.

[17] G. Lancia, V. Bafna, S. Istrail, R. Lippert, and R. Schwartz, "Snps problems, complexity, and algorithms," in *Algorithmsesa*. Springer Berlin Heidelberg, 2001, pp. 182–193.

[18] R. S. Wang, L. Y. Wu, Z. P. Li, and X. S. Zhang, "Haplotype reconstruction from snp fragments by minimum error correction," in *Bioinformatics*, 2005, vol. 21(10), pp. 2456–2462.

[19] S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, et al., "The diploid genome sequence of an individual human," in *PLoS Biol*, 2005, vol. 5(10), p. e254.

[20] V. Bansal and V. Bafna, "Hapcut: an efficient and accurate algorithm for the haplotype assembly problem," in *Bioinformatics*, 2008, vol. 24(16), pp. i153–i159.

[21] V. Bansal, A. L. Halpern, N. Axelrod, and V. Bafna, "An mcmc algorithm for haplotype assembly from whole-genome sequence data," in *Genome research*, 2008, vol. 18(8), pp. 1336–1346.

[22] J. Duitama, G. K. McEwen, T. Huebsch, S. Palczewski, S. Schulz, K. Verstrepen, et al., "Fosmid-based whole genome haplotyping of a hapmap trio child: evaluation of single individual haplotyping techniques," in *Nucleic acids research*, 2011, p. gkr1042.

[23] C. Cai, S. Sanghavi, and H. Vikalo, "Structured low-rank matrix factorization for haplotype assembly," in *Journal of Selected Topics in Signal Processing*. IEEE, 2016, vol. 10.4, pp. 647–657.

[24] S. Gunasekar, A. Acharya, N. Gaur, and J. Ghosh, "Noisy matrix completion using alternating minimization," in *ECML PKDD*. Springer, 2013, pp. 194–209.

[25] R.H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," in *Journal of Machine Learning Research*, 2010, vol. 11, pp. 2057–2078.

[26] V. Kuleshov, "Probabilistic single-individual haplotyping," in *Bioinformatics*, 2014, vol. 30(17), pp. i379–i385.

[27] H. Matsumoto and H. Kiryu, "Mixsih: a mixture model for single individual haplotyping.," in *BMC Genomics*, 2013, vol. 14(Suppl 2) S5.

[28] F. Geraci, "A comparison of several algorithms for the single individual snp haplotyping reconstruction problem," in *Bioinformatics*, 2010, vol. 26(18), pp. 2217–2225.

[29] X. Chen, Q. Peng, L. Han, T. Zhong, and T. Xu, "An effective haplotype assembly algorithm based on hypergraph partitioning," in *Journal of theoretical biology*, 2014, vol. 358, pp. 85–92.