Denoising individual bias for a fairer binary submatrix detection

Changlin Wan
Purdue University
wan82@purdue.edu

Wennan Chang
Purdue University
chang534@purdue.edu

Tong Zhao
Amazon
zhaoton@amazon.com

Sha Cao
Indiana University
shacao@iu.edu

Chi Zhang
Indiana University
czhang87@iu.edu

ABSTRACT

Low rank representation of binary matrix is powerful in disentangling sparse individual-attribute associations, and has received wide applications. Existing binary matrix factorization (BMF) or co-clustering (CC) methods often assume i.i.d background noise. However, this assumption could be easily violated in real data, where heterogeneous row- or column-wise probability of binary entries results in disparate element-wise background distribution, and paralyzes the rationality of existing methods. We propose a binary data denoising framework, namely BIND, which optimizes the detection of true patterns by estimating the row- or column-wise mixture distribution of patterns and disparate background, and eliminating the binary attributes that are more likely from the background. BIND is supported by thoroughly derived mathematical property of the row- and column-wise mixture distributions. Our experiment on synthetic and real-world data demonstrated BIND effectively removes background noise and drastically increases the fairness and accuracy of state-of-the arts BMF and CC methods.

CCS CONCEPTS

• **Computing methodologies** → **Representation of mathematical objects**; *Representation of Boolean functions.*

KEYWORDS

Binary data mining, fairness, denoising, low rank representation

ACM Reference Format:
Changlin Wan, Wennan Chang, Tong Zhao, Sha Cao, and Chi Zhang. 2020. Denoising individual bias for a fairer binary submatrix detection. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3340531.3412156

1 MOTIVATION

Binary matrix has been commonly utilized in multiple fields. Low rank pattern in a binary matrix is defined as rank-1 sub matrices formed by the product of two binary bases. Comparing to continuous data, recent studies demonstrated the rank-1 sub-matrices in binarized data is more robust for mechanism interpretation or subspace representation [3, 4], because binary data in general bears reduced noise than continuous data. However, variations of the probability of 1s of rows or columns may lead to varied element-wise probability, causing a fairness issue in low rank representation of binary data [6].

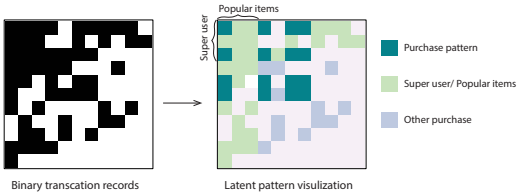

Figure 1: Individual bias in binary transaction records data

An intuitive example is binary transaction records data (figure 1), in which 1s represent the purchase of items (each column) by users (each row). Different items or users are with varied activities in conducting purchasing. For example, super-users make more purchase, which can be independent to items, and popular items are more likely to be purchased. The transactions made between super users and popular items unnecessarily imply good recommendations since it can be simply caused by the high purchase chance. On the other hand, the group of items having a strong purchase preference within a certain group of users comparing to their background purchase rate is more valuable for recommendation. However, the fairness issue in the low rank representation of binary data due to varied element-wise background probability was rarely considered in existing formulations [5].

Here, we propose BIND, a binary data denoising method via considering the data is generated from the mixture of to-be-identified rank-1 patterns and an unknown background of element-wise probability, plus i.i.d. errors. BIND estimates the mixture distribution of the probabilities of 1s from rank-1 patterns and background in each row and column, by which the rows or columns that are more likely with true rank-1 patterns are distinguished by the over-represented 1s comparing to the background.

Key contributions of this work include: (1) BIND is the first of this kind of binary data denoising method via considering non-identical background distribution, (2) BIND can be easily implemented with state-of-the-arts BMF or CC methods for a fairer rank-1 pattern detection, and (3) rigorous mathematical derivations are provided to characterize the property of disparate background distribution.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412156

2 BACKGROUND

2.1 Notations

We denote matrix, vector and scalar by uppercase, bold lowercase and lowercase character $X, \mathbf{x}, x$. Superscript with $\times$ indicates dimensions, while subscript implies index, such as $X_{ij}^{m\times n}$ and $\mathbf{x}_i^{m\times 1}$. $P_{ij} \triangleq P(X_{ij} = 1)$ denotes the element-wise probability of 1 at the element X_{ij}. $|\mathbf{x}|$ and $|X|$ represent the $l1$ norm of vector and matrix, and $\circ$ represents Hadamard product.

2.2 Related work

Existing methods of binary matrix low rank representation fall into two major categories, namely binary matrix decomposition (BMF) and co-clustering (CC). BMF aims to decompose a binary matrix as the product of two low rank binary matrix by maximizing its overall fitting to the original matrix. The formulation of BMF is thus generalized as

$$X^{m\times n} = U^{m\times k} V^{k\times n} + E^{m\times n}$$

, where U and V are the low rank pattern matrices, and E is the flipping error with $p(1 \to 0) = p(0 \to 1) = p_0$. BMF problem is NP-hard, for which multiple heuristic algorithms have been developed. One representative method is ASSO, which retrieves candidate patterns by using row-/column-wise correlation [2]. More recently, Bayesian probability measure and geometrical identification largely improved the efficiency and accuracy of BMF [3, 4].

In contrast, the co-clustering (CC) method, also named as bi-clustering in statistics and computational biology, maximizes the enrichment of 1s in the detected patterns based on certain thresholds[1]. For given $X^{m\times n}$, most CC methods aim to identify the cardinality of index set $I_l \times J_l, l = 1, ..., k$, where $I_l \in \{1, ..., m\}$ and $J_l \in \{1, ..., n\}$,

$$s.t. \quad P_{ij} = \begin{cases} p_l, \ if \ i, j \in I_l \times J_l \\ p_0, \ if \ i, j \notin I_l \times J_l \end{cases} \quad \forall l = 1, ..., k$$

Noted, both BMF and CC methods assume the binary data is formed by the sum of to-be-identified rank-1 submatrices and an i.i.d error, where individuals bias has not been investigated.

2.3 Problem formulation

We consider the observed binary data with disparate element-wise background probability that is generated by:

$$X = U^{m\times k} V^{k\times n} + X^0 + E' + E \qquad (\star)$$

Compared with the formulation of BMF, X^0 is the background matrix. E' is the pattern wise observation error that each element from pattern l has a probability of $1 - p_l$ to be zero, while the elements outside patterns will not be impacted, i.e., $P_{ij}^{E'}(1 \to 0) = 1 - p_l$, if $i, j \in I_l \times J_l$, $P_{ij}^{E'}(1 \to 0) = 0$, if $i, j \notin I_l \times J_l$, $\forall l = 1, ..., k$.

Under this definition, by considering X^0 are 0, current BMF and CC described in 2.2 are special case of $(\star)$, and were designed to handle the pattern observation error E' and elment-wise flipping error E. Thus, the bottleneck of a fair binary submatrix detection lies in differentiating true patterns from the background X^0. We consider the assumption of $P(X_{ij}^0 = 1) \propto \mathbf{p}_i^{0,r} \cdot \mathbf{p}_j^{0,c}$ that can cover most of the binary data with disparate background, when X_{ij}^0 are conditionally independent with fixed row or column index, like the purchase transaction data in figure 1 with items of different popularity and users of different activity. We denote the row/column-wise background probability as $\mathbf{p}^{m\times 1,\ 0,row}$ and $\mathbf{p}^{n\times 1,\ 0,column}$, shorted as $\mathbf{p}^{0,r}$ and $\mathbf{p}^{0,c}$, where $\mathbf{p}_i^{0,r} \propto \hat{\mathbf{p}}_i^{0,r} = \frac{|X_{i:}^0|}{n}$ and $\mathbf{p}_j^{0,c} \propto \hat{\mathbf{p}}_j^{0,c} = \frac{|X_{:j}^0|}{m}$, and $P(X_{ij}^0 = 1)$ can be unbiasedly estimated as $\frac{|X_{i:}^0| \cdot |X_{:j}^0|}{|X^0|}$.

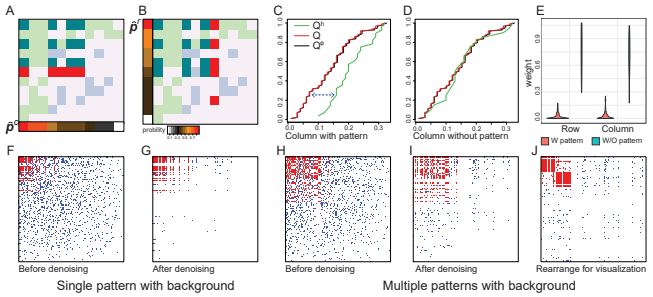

Figure 2: quantile shift denoising

3 BIND FRAMEWORK

Here we propose the BIND[1] framework to identify the rank-1 patterns (U, V) from binary data X with disparate background X^0. Denoting $P(X_{ij}^0 = 1)$ as P_{ij}^0, the element-wise probability $P_{ij} \triangleq P(X_{ij} = 1)$ can be derived as:

$$P_{ij} = \begin{cases} P_{ij}^0 \propto \mathbf{p}_i^{0,r} \cdot \mathbf{p}_j^{0,c}, \ ij \notin any \ I_l \times J_l \\ 1 - (1 - P_{ij}^0)(1 - p_l) = p_{ij}^0 + (1 - p_{ij}^0)p_l, \ ij \in I_l \times J_l \end{cases} \qquad (*)$$

Specifically, the row and column probability $\mathbf{p}_i^r$ and $\mathbf{p}_j^c$ can be estimated by $\hat{\mathbf{p}}_i^r = \frac{|X_{i:}|}{n}$ and $\hat{\mathbf{p}}_j^c = \frac{|X_{:j}|}{m}$. Noted, $\mathbf{p}^r$ and $\mathbf{p}^c$ are formed by the mixture distribution of $\mathbf{p}^{0,r}, \mathbf{p}^{0,c}$ and p_l. Analogous to BMF and CC problem, direct inference of $\mathbf{p}^{0,r}, \mathbf{p}^{0,c}$ and p_l from $\mathbf{p}^r$ and $\mathbf{p}^c$ is NP-hard. As shown in Figure 2A-D, instead of computing $\mathbf{p}^{0,r}, \mathbf{p}^{0,c}$ and p_l, BIND identifies the rows and columns that are most likely conceiving patterns comparing to others. The elements of the intersection of the identified rows and columns more likely represent true rank-1 patterns (figure 2F-J). For this task, we introduce the quantile_shift algorithm with thorough mathematical proof.

Quantile_shift algorithm is designed to distinguish rows or columns that are more likely conceiving rank-1 patterns. First, we introduce the concept of empirical distribution of row-/column-wise probability, denoted as $\mathbf{F}^r$ and $\mathbf{F}^c$ (figure 2A,B), which are sampled from $\hat{\mathbf{p}}^r$ and $\hat{\mathbf{p}}^c$ with probability $P(\mathbf{F}^r = \hat{\mathbf{p}}_i^r) \propto \hat{\mathbf{p}}_i^r$ and $P(\mathbf{F}^c = \hat{\mathbf{p}}_j^c) \propto \hat{\mathbf{p}}_j^c$. The observed probability of hits $\mathbf{F}^h$ of any row i_0 or column j_0 is defined by $\mathbf{F}^{h,r,i_0} = \{\hat{\mathbf{p}}_j^c | j \ with \ X_{i_0 j} = 1\}$ and $\mathbf{F}^{h,c,j_0} = \{\hat{\mathbf{p}}_i^r | i \ with \ X_{i j_0} = 1\}$. Here $\mathbf{F}^r$ and $\mathbf{F}^c$ characterize the distribution of $\hat{\mathbf{p}}^r$ and $\hat{\mathbf{p}}^c$ of the 1s randomly drawn from $\hat{\mathbf{p}}^r$ and $\hat{\mathbf{p}}^c$. Intuitively, if a row or column conceives a distinct pattern, the quantile function Q^h of $\mathbf{F}^h$ will shift drastically from the quantile function Q^c of $\mathbf{F}^c$ or Q^r of $\mathbf{F}^r$ (figure 2C). On the other hand, Q^h will be similar to Q^c or Q^r if the row or column does not contain

[1]Code and material can be access at https://github.com/clwan/BIND

any pattern (figure 2D). Hence the shift between Q^h and Q^r or Q^c can serve as a weight s to differentiate the rows or columns more likely conceiving a pattern (figure 2E). Noted, here $\mathbf{F}^r$ and $\mathbf{F}^c$ serve as proxy of $\mathbf{F}^{0,r}$ and $\mathbf{F}^{0,c}$, which are the empirical distribution of the true background probability of $\mathbf{p}^{0,r}$ and $\mathbf{p}^{0,c}$. In the following content, we prove s approximates the pattern size within each row or column, i.e., $s \approx |(UV+E')_{i:}|$ *or* $|(UV+E')_{:j}|$ with certain bounds.

The input of *Quantile_shift* algorithm include a row or column index i_0/j_0, and $\hat{\mathbf{p}}^c$ or $\hat{\mathbf{p}}^r$, by which the empirical distribution $\mathbf{F}^c$ or $\mathbf{F}^r$ will be sampled, and the probability of hit of the row or column $\mathbf{F}^h$ will be computed. The output is weight s of the row or column. Without loss of generality, we illustrate the *Quantile_shift* algorithm for computing the weight of row i_0 below, and detailed mathematical proofs as follows:

Algorithm 1: Quantile_shift

Inputs: Row index i_0, Estimated column-wise probability $\hat{\mathbf{p}}^c$
Outputs: Estimated weight of significance of row i_0, $s^r_{i_0}$
Quantile_shift$(i_0, \hat{\mathbf{p}}^c)$:
$\mathbf{F}^c \leftarrow sampled\ from\ \hat{\mathbf{p}}^c\ with\ probability\ \hat{\mathbf{p}}^c$
$\mathbf{F}^h \leftarrow \{\hat{\mathbf{p}}^c_j | j\ with\ X_{i_0 j} = 1\}$
$\mathbf{F}^{(h)} \leftarrow sort(\mathbf{F}^h)$, $a \leftarrow length(\mathbf{F}^h)$
$Q^c(p) = sup(b)\ s.t.\ \frac{|\mathbf{F}^c < b|}{length(\mathbf{F}^c)} \le p\ and\ \frac{|\mathbf{F}^c > b|+1}{length(\mathbf{F}^c)} > p$
for *j=1...a* **do**
  if $F^{(h)}_j > Q^c(\frac{j}{a})$ **then**
    $t_j \leftarrow$ *the column index s.t.* $\mathbf{F}^{(h)}_j = \hat{\mathbf{p}}^c_{t_j}$ & $X_{i_0 t_j} = 1$
    $s \leftarrow s + \frac{\mathbf{F}^{(h)}_j - Q^c(\frac{j}{a})}{1-\hat{\mathbf{p}}^c_{t_j}}$
end

Lemma 1. *If $\hat{\mathbf{p}}^r$ and $\hat{\mathbf{p}}^c$ are unbiased estimation of $\mathbf{p}^{0,r}$ and $\mathbf{p}^{0,c}$. The weight computed by quantile_shift is an unbiased estimation of the sum of $E(U^{m\times k}V^{k\times n} + E')$ with respect to that column or row.*

Proof. If $\hat{\mathbf{p}}^r$ and $\hat{\mathbf{p}}^c$ are unbiased estimation of $\mathbf{p}^{0,r}$ and $\mathbf{p}^{0,c}$, $\mathbf{F}^r$ or $\mathbf{F}^c$ generated from $\hat{\mathbf{p}}^r$ and $\hat{\mathbf{p}}^c$ form unbiased empirical distribution of row-/column-wise probability of 1s of X^0, i.e. $P(\mathbf{F}^{0,r} = \mathbf{p}^{0,r}_i) \propto \mathbf{p}^{0,r}_i$ and $P(\mathbf{F}^{0,c} = \mathbf{p}^{0,c}_j) \propto \mathbf{p}^{0,c}_j$. Without loss of generality, we prove the lemma for the computation of the weight of the i_0th row. Denote $\mathbf{t} = \{j | X_{i_0 j} = 1\}$ and $a = length(\mathbf{t})$, by **Algorithm 1** and $(*)$, $\forall\ j \in \{1, ..., a\}$:
If $i_0 t_j \notin any\ I_l \times J_l$,

$$E(\mathbf{F}^{(h)}_j - Q(\frac{j}{a})) = E(\hat{\mathbf{p}}^c_{\mathbf{t}_j} - sup(b|\frac{|\mathbf{F}^c < b|}{length(\mathbf{F}^c)} \le \frac{j}{a})) = 0$$

Else, $i_0 t_j \in I_l \times J_l\ for\ certain\ l$,

$$E(\mathbf{F}^{(h)}_j - Q(\frac{j}{a})) = E(\hat{\mathbf{p}}^c_{\mathbf{t}_j} + (1-\hat{\mathbf{p}}^c_{\mathbf{t}_j})p_l - sup(b|\frac{|\mathbf{F}^c < b|}{length(\mathbf{F}^c)} \le \frac{j}{a})) = (1-\hat{\mathbf{p}}^c_{\mathbf{t}_j})p_l$$

Such that

$$E(\sum_{j=1}^{a}\frac{\mathbf{F}^{(h)}_j - Q(\frac{j}{a})}{1-\hat{\mathbf{p}}^c_{\mathbf{t}_j}}) = \sum_l\sum_{j=1}^{a} p_l I = |E(U^{m\times k}V^{k\times n} + E')_{i_0:}|$$

□

Lemma 2. *For X in $(\star)$, and $P^0_{ij} \triangleq P(X^0_{ij}) \propto \boldsymbol{p}^{0,r}_i \cdot \boldsymbol{p}^{0,c}_j$, the probability estimated by $\hat{p}^r_i = \frac{|X_{i:}|}{n}$ and $\hat{p}^c_j = \frac{|X_{:j}|}{m}$ are bounded by $|\hat{p}^r_i - \boldsymbol{p}^{0,r}_i| \le \frac{\sum_{l=1}^{k}\mathbf{1}(i\in I_l)p_l|J_l|}{n}$, and $|\hat{p}^c_j - \boldsymbol{p}^{0,c}_j| \le \frac{\sum_{l=1}^{k}\mathbf{1}(j\in J_l)p_l|I_l|}{m}$.*

Lemma 2 can be derictly derived from $(\star)$ and $(*)$.

Lemma 3. *The weight of the i_0th row (or similarly j_0th column) is with a bias led by the biasedly estimated $\hat{\boldsymbol{p}}^c$ and $\hat{\boldsymbol{p}}^r$, which is bounded by $E(s - |(UV+E')_{i_0:}|) \le \frac{max(\boldsymbol{F}^c)+max(\frac{|E(UV+E')_{:j}|}{m})(|\boldsymbol{F}^c|+1)}{min(1-\boldsymbol{p}^h)|\boldsymbol{F}^c|}$.*

We still use the compututaion of the i_0th row to illustrate the proof. The case for columns can be similarly derived.

Proof. By Lemma 2, $\hat{\mathbf{p}}^c$ is a biased estimation of $\mathbf{p}^{0,c}$, where $\hat{\mathbf{p}}^c_j = \frac{|X_{:j}|}{m} \ge \mathbf{p}^{0,c}_j = \frac{|X^0_{:j}|}{m}$, $j = 1, ..., m$. Hence $\mathbf{F}^{(h)} \ge \mathbf{F}^{0,(h)}$, suggesting $1 - \mathbf{F}^{0,(h)} \ge 1 - \mathbf{F}^{(h)}$ and $Q^c(\frac{j}{a}) \ge Q^{0,c}(\frac{j}{a})$, by which

$$\left|\frac{\mathbf{F}^{0,(h)}_j - Q^{0,c}(\frac{j}{a})}{1-\hat{\mathbf{p}}^{0,c}_{\mathbf{t}_j}} - \frac{\mathbf{F}^{(h)}_j - Q^c(\frac{j}{a})}{1-\hat{\mathbf{p}}^c_{\mathbf{t}_j}}\right| \le 2\left|\frac{\max_{z\in(0,1)}\{Q^c(z) - Q^{0,c}(z)\}}{1-\hat{\mathbf{p}}^c_{\mathbf{t}_j}}\right|$$

By lemma 2, the bias of $|\hat{\mathbf{p}}^c_j - \hat{\mathbf{p}}^{c,0}_j|$ is bounded by $\frac{|E(UV+E')_{:j}|}{m}$. So the max shift caused in the quantile function $\max_{z\in(0,1)}\{Q^c(z) - Q^{0,c}(z)\}$ is bounded by $\frac{max(\hat{\mathbf{p}}^c)+max(\frac{|E(UV+E')_{:j}|}{m})}{|\hat{\mathbf{p}}^c|}+max(\frac{|E(UV+E')_{:j}|}{m})$. Hence the cumulative bias is bounded by

$$E(s-|E(UV+E')_{i_0:}|) \le \frac{a(max(\hat{\mathbf{p}}^c) + max(\frac{|E(UV+E')_{:j}|}{m})(|\hat{\mathbf{p}}^c|+1)))}{min(1-\hat{\mathbf{p}}^c)|\hat{\mathbf{p}}^c|}$$

□

Lemma 1 suggests $|Q^h$-$Q^0|$ is an unbiased estimation of the expected number of 1s in the rank-1 patterns and Lemma 2-3 provide the bound of the bias of $|Q^h$-$Q|$ when Q^0 is biasedly estimated as Q.

Theorem 1 (Quantile_shift). *For a relative sparse binary matrix, the weight calculated by Quantile_shift sufficiently characterizes the indices of the patterns with largest $P_l|I_l|$ and $P_l|J_l|$.*

Proof. For i0th row (or similarly for the j0th column),

$$E(s - |(UV+E')_{i_0:}|) \le \frac{a(max(\hat{\mathbf{p}}^c) + max(\frac{|E(UV+E')_{:j}|}{m})(|\hat{\mathbf{p}}^c|+1)))}{min(1-\hat{\mathbf{p}}^c)|\hat{\mathbf{p}}^c|}$$
$$\approx \frac{a}{min(1-\hat{\mathbf{p}}^c)}max\{\frac{max(\hat{\mathbf{p}}^c)}{|\hat{\mathbf{p}}^c|}, max(\frac{|E(UV+E')_{:j}|}{m})\}$$

, suggests that when the input matrix and rank-1 patterns are relatively sparse, the weight s approximates $(UV+E)_{i_0:}$, i.e. largest values in $\mathbf{s}^r$ and $\mathbf{s}^c$ correspond to the rows and columns of the patterns with largest $P_l|I_l|$ and $P_l|J_l|$. □

BIND framework is developed to implement *Quantile_shift* algorithm with a BMF or CC method, denoted as $\mathcal{F}$, for a fairer rank-1 pattern identification under the formulation of $(\star)$. As illustrated in figure 2F-J, *Quantile_shift* denoises the majority of the background signal and enables a BMF or CC method better detects $U^{m\times k}$ and $V^{k\times n}$. A cutoff τ is needed to differentiated the weight

of the rows or columns with true patterns (figure 2E). Empirically, τ could be set from 0.05 to 0.1 in BIND algorithm.

BIND is capable for one direction denoising. The $Quantile\_shift$ algorithm is $O(n)$ or $O(m)$ for row or column weight computation and the BIND algorithm is $O(mn)$, which is smaller than most of current BMF and CC methods. The BIND algorithm is detailed below:

Algorithm 2: BIND

Inputs: Input data $X^{m\times n}$, Threshold τ, BMF/CC method $\mathcal{F}$
Outputs: Pattern matrices $U^{m\times k}$ and $V^{k\times n}$
$BIND(X, \tau, \mathcal{F})$:
$X_{use} \leftarrow 0 \cdot X$, $\mathbf{s}^r \leftarrow \mathbf{0}^{m\times 1}$, $\mathbf{s}^c \leftarrow \mathbf{0}^{n\times 1}$
$\hat{\mathbf{p}}_i^r = \frac{|X_{i:}|}{n}\ \forall i = 1, ..., m$ and $\hat{\mathbf{p}}_j^c = \frac{|X_{:j}|}{m}\ \forall j = 1, ..., n$
for *i=1...m* **do**
  $\mathbf{s}_i^r = Quantile\_shift(i, \hat{\mathbf{p}}^c)$
end
for *j=1...n* **do**
  $\mathbf{s}_j^c = Quantile\_shift(j, \hat{\mathbf{p}}^r)$
end
$I^r \leftarrow I(\mathbf{s}^r > \tau), I^c \leftarrow I(\mathbf{s}^c > \tau), X_{use} \leftarrow X \circ (I^r \cdot I^{cT})$
$U, V \leftarrow \mathcal{F}(X_{use}, ...)$

4 EXPERIMENT

In this section, we evaluate the performance of BIND on synthetic and real-world data sets across different data scenarios. We demonstrate the implementation of BIND with different BIND BMF and CC methods can significantly improve their fairness in detecting rank-1 pattern from binary matrix with disparate background probability. We also highlight the application of BIND framework for better result interpretation on real-world Movielens data.

We simulate synthetic data sets $X^{100\times 100}$ with fixed size by following ($\star$): $X = U^{m\times k}V^{k\times n} + E' + X^0 + E$, with different pattern size $\in \{10, 15, 20\}$, pattern number $k \in \{1, 2\}$, observation error $p_k \in \{0.8, 0.9, 1.0\}$, background probability $\mathbf{p}^{0,r}, \mathbf{p}^{0,c}$, and element-wise flipping error $p_0 \in \{0, 0.05\}$. Specifically, background probabilities were generated from uniform distribution $\mathbf{p}^{0,r}, \mathbf{p}^{0,c} \sim U[0.1, p]$, where $p \in \{0.5, 0.6, 0.7\}$ corresponds to different background probabilities. Altogether, we deem 108 data scenarios from the above parameter settings and simulated 30 replicates for each scenario to form a test-bed. Jaccard index $\mathbf{D} = \frac{|X\cap UV|}{|X\cup UV|}$ ($X =$ *original or denoised data*) is used as the evaluation metric. For each data scenario, denoising performance is evaluated by the averaged Jaccard index on the 30 replicates. We first compare the performance with respect to different significance threshold $\tau = \{0, 0.05-1\}$, where $\tau = 0$ represents the data without denoising. As shown in figure 3A, the denoising process on average increased the Jaccard index by 2.6 fold and denoising efficiency is slightly increased with τ. Table 1 lists the denoising performance with respect to different number of patterns k, background probability$p$ and observation probability p_k, where pattern size is set as 15 and $\tau = 0.1$.

We benchmark BIND by implementing with recently developed BMF method LOM and CC method Biclust, which showed top performance among similar state-of-the-arts methods [1, 3]. The implementation of BIND largely increased the accuracy in detecting true patterns, which results in an averaged 7.5 (LOM) and 2.6 (Biclust) fold increase of the Jaccard index (figure 3B,C) .

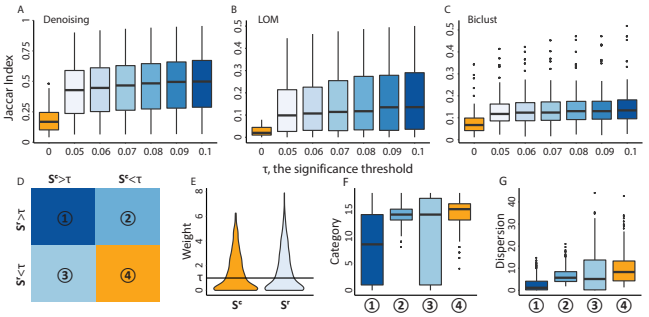

Figure 3: Performance on simulated and Movielens data

p \ p_k	single pattern			Multiple pattern		
	0.8	0.9	1.0	0.8	0.9	1.0
0.5	0.17/0.67	0.18/0.79	0.20/0.88	0.28/0.59	0.31/0.73	0.34/0.84
0.6	0.13/0.48	0.14/0.61	0.16/0.73	0.23/0.47	0.26/0.59	0.28/0.69
0.7	0.11/0.29	0.11/0.37	0.13/0.47	0.19/0.34	0.21/0.40	0.22/0.52

Table 1: Jaccard index before/after denoising

We also demostrate that BIND increases the interpretation and denoising in real-world Movielens data, in which $X_{ij} = 1$ represents the interest of user i (row) in rating/watching movie j (column). Category label of each movie is provided. Intuitively, disparate background probablities naturally exist in this data due to different popularity of movies and activity of users. Data is divided into four regions by the I^r and I^c computed in **Algorithm 2** (figure 3D,E), where ① is the region most likely with patterns, and ②, ③ and ④ are denoised regions. Users in region ① watched more movies but less categories comparing to other regions (figure 3F), suggesting potential recommendation. In addition, region ① has smallest dispersion of the number of rated movies with respect to different categories, suggesting more stable rating preference of users towards their preferred movie types in this region (figure 3G).

5 ACKNOWLEDGMENTS

This work was supported by R01 award #1R01GM131399- 01, NSF IIS (N0.1850360), Showalter Young Investigator Award from Indiana CTSI and Indiana University Grand Challenge Precision Health Initiative.

REFERENCES

[1] Sebastian Kaiser and Friedrich Leisch. 2008. A toolbox for bicluster analysis in R. (2008).
[2] Pauli Miettinen, Taneli Mielikäinen, Aristides Gionis, Gautam Das, and Heikki Mannila. 2008. The discrete basis problem. *IEEE transactions on knowledge and data engineering* 20, 10 (2008), 1348–1362.
[3] Tammo Rukat, Chris C Holmes, Michalis K Titsias, and Christopher Yau. 2017. Bayesian boolean matrix factorisation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2969–2978.
[4] Changlin Wan, Wennan Chang, Tong Zhao, Mengya Li, Sha Cao, and Chi Zhang. 2019. Fast and efficient Boolean matrix factorization by geometric segmentation.

arXiv:1909.03991 (2019).

[5] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*. 2921–2930.

[6] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM CIKM*. 1153–1162.