IISE Transactions on Healthcare Systems Engineering

ISSN: 2472-5579 (Print) 2472-5587 (Online) Journal homepage: https://www.tandfonline.com/loi/uhse21

Missing information imputation for disease-dedicated social networks with heterogeneous auxiliary data

Xu Liu, Jingrui He, Wanli Min & Hongxia Yang

To cite this article: Xu Liu, Jingrui He, Wanli Min & Hongxia Yang (2020) Missing information imputation for disease-dedicated social networks with heterogeneous auxiliary data, IISE Transactions on Healthcare Systems Engineering, 10:2, 87-98, DOI: 10.1080/24725579.2020.1716115

To link to this article: https://doi.org/10.1080/24725579.2020.1716115

Accepted author version posted online: 15 Jan 2020.
Published online: 30 Jan 2020.

IISE TRANSACTIONS ON HEALTHCARE SYSTEMS ENGINEERING
2020, VOL. 10, NO. 2, 87–98
https://doi.org/10.1080/24725579.2020.1716115

Missing information imputation for disease-dedicated social networks with heterogeneous auxiliary data

Xu Liu[a], Jingrui He[a], Wanli Min[b], and Hongxia Yang[b]

[a]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA; [b]Alibaba Group, Hangzhou, China

ABSTRACT

Many high impact applications suffer from missing information. For example, disease-dedicated social networks provide additional resources to glimpse into patients' daily life related to disease management. However, due to the voluntary nature of such social networks, the information reported by patients is often incomplete, making the following data analytics tasks particularly challenging. On the other hand, in addition to the target data that we aim to analyze, we may also have other related data at our disposal. For example, to analyze disease-dedicated social networks, auxiliary clinical data (with potentially non-overlapping patients), as well as the users' online social relationship might provide additional information for estimating the missing information. Therefore, the key question we aim to answer in this paper is how we can leverage the heterogeneous auxiliary data for the sake of missing information imputation. To answer this question, we focus on diabetes-dedicated social networks, and we aim to estimate the missing information from patients' self-reported biomarker measurements. In particular, we propose a hypergraph structure to model the relationship among users and user-generated content (posts). Based on the hypergraph structure, we further introduce an optimization framework to estimate the missing biomarker measurements using heterogeneous auxiliary data. To solve the optimization framework, we design iterative algorithms to find the local optimal solution. Experimental results on both synthetic and real data sets (including a data set collected from a diabetes-dedicated social network) demonstrate the effectiveness of the proposed algorithms.

ARTICLE HISTORY
Received 28 August 2018
Accepted 31 December 2019

KEYWORDS
Disease-dedicated social network; heterogeneous learning; missing value imputation

1. Introduction

We are living in as the era of big data, despite a large amount of data being collected across multiple areas, a commonly observed phenomenon is missing information and missing value. For example, recent years have witnessed the rapid growth of disease-dedicated social networks. Different from the generic social networks, the disease-dedicated social networks are designed for and used by patients with the same type of disease, such as diabetes mellitus. Their goal is to enable information sharing and support group formation, which in turn, can help patients maintain a healthy lifestyle concerning the disease. On the disease-dedicated social networks, although users often report their biomarker measurements as their condition progresses, such self-reported measurements contain a large amount of missing information, as very few users report their measurements every time they take the test. However, for monitoring purposes, it is essential to have a reliable estimate of such missing information, so that reminders or alerts can be generated in time to help users get back on track.

On the other hand, in order to estimate the missing information and improve the imputation performance, we often have access to heterogeneous auxiliary data in addition to the observed information. For example, to estimate the missing biomarker measurements from disease-dedicated social networks, in addition to the self-reported measurements, we can also leverage the rich social relations exist in a large amount of the frequent visitor, users, such as friend–friend relationship, follower–followee relationship. Furthermore, beyond the disease-dedicated social networks, auxiliary clinical data with potentially non-overlapping users may reveal an important trend regarding the progression of the biomarker measurements, and thus it can help improve the performance of missing value imputation.

In this paper, motivated by missing biomarker measurement estimation on disease-dedicated social networks, we propose a generic framework for **M**issing **I**nformation **I**mputation with **H**ypergraph structure using **D**ual matrix factorization named MI^2-HD, or **T**riple matrix factorization named MI^2-HT. It effectively leverages both the observed values and additional information from heterogeneous sources. In particular, we propose to use a hypergraph structure to model the relationship among users and user-generated content, or posts; and we design an optimization problem based on joint matrix factorization (MF), which uses both the rich social relationship among the users as well as

CONTACT Jingrui He ✉ jingrui.he@gmail.com School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uhse.

auxiliary clinical data to improve the accuracy of the missing value imputation. The performance of the resulting iterative algorithms is demonstrated on both synthetic and multiple real data sets.

The main contributions of this paper can be summarized as follows:

- **Problem setting**. We propose a novel problem setting of missing value imputation for disease-dedicated social networks. The missing value of the patients' medical and healthcare measurement is estimated in accordance with the auxiliary data and the disease-specific grouping pattern enforced by the hypergraph structure.
- **Optimization framework**. The proposed problem setting is formulated as an optimization framework based on joint MF. We further introduce efficient multiplicative update rules to solve this framework.
- **Experimental evaluation**. Multiple experiments have been conducted on one synthetic data set, one benchmark data set, and one real data set collected from a disease-dedicated social network. Experimental results demonstrate the effectiveness of the proposed algorithms.

The rest of the paper is organized as follows. In Section 2, we review the related work about missing value imputation and MF. The proposed missing value imputation framework is introduced in Section 3. Section 4 presents the iterative optimization algorithms, followed by the analysis of their convergence and time complexity. Section 5 shows the experimental results, and finally, we conclude the paper in Section 6.

2. Related work

In this section, we briefly review the related works on missing value imputation and MF.

2.1. Missing value imputation

The missing data is evolving as a critical issue in the field of medical and healthcare data analysis. The missing data occurs due to various reasons in the medical study, such as the lack of collection, and the lack of documentation (Wells, Chagin, Nowacki, & Kattan, 2013). Lack of collection usually happens in the datum of the disease-dedicated social network as the consequence of the voluntary nature of the online user. The lack of documentation, such as the electronic health record (EHR), which is designed to the benefit of clinical and billing company, is particularly prevailing when the clinical measurement shows the negative symptom/comorbidity thus all the record fields are left blank instead of recording the values. Overall, the issue of missing value is commonly present in various types of medical and healthcare data. In general, there are three types of missing mechanisms, i.e. missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), disturb the healthcare data analysis (Pedersen et al., 2017). Each of the mechanism requires different analysis methods due to their own characteristics as:

- MCAR: In this case, the missing entries occur at completely random as the value of the missing entries has no dependence on the observed knowledge. Any kind of the data imputation method can be adopted without bringing in the bias risk as no previous constraint specification (Janssen et al., 2010).
- MAR: Compared with the MCAR, in which no specific constraint exists between the missing data and observed data, MAR means the observed variables can partially explain the missing data. For example, when the blood pressure data is MAR, the variables of age and gender are considered as the dependent variables to the blood pressure, compared with the variable of estrogen receptor (ER) or progesterone receptor (PR) that indicate the breast cancer statue.
- MNAR: In this type of missing mechanism, the unobserved variables are assumed to be related to the values of that variable itself, i.e. the missing value is specifically related to what is missing. An example of MNAR in disease-dedicated social network analysis occurs if those heavy patients may be less likely to disclose their weight.

As the amount of medical and healthcare data presence the missing value issue, the need of missing value imputation is growing in order to provide the comprehensive research data, instead of ignoring the entire row or column where missing value happens (Janssen et al., 2010). Previous missing value imputation methods handle the medical data by two strategies: (1) Single value imputation (SVI) and (2) multiple-imputation (MI). Both of these two strategies are mainly focused on the MCAR and MAR missing mechanism, as the MNAR leads to the most difficult case when none assumption has been made about what is happening in the missing data. SVI aims at estimating the unknown entries by a single value. For example, the most commonly used method is to replace the missing entries by the overall mean value of the observed entries (Donders, Van Der Heijden, Stijnen, & Moons, 2006), or using the most commonly observed values to recover the missing entries (Luengo, García, & Herrera, 2012). Another widely used method is regression imputation (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001) (also known as the predicted mean imputation). They are straightforward to understand, but tend to underestimate the diversity of the original data, and also ignoring the correlations between the samples. The SVI strategy is leading to the biased imputation result and causes the Type 1 error (i.e. the none existing relation is identified) (Greenland & Finkle, 1995), which may not be suitable for many real-world applications (Enders, 2010). Compared with SVI, the MI strategy aims at predicting the missing entries value based on the distribution of observed knowledge, such as the expectation-maximization (EM) based method (Musil, Warner, Yobas, & Jones, 2002), MF-based method (Lange & Buhmann, 2006). The MF-based

Table 1. Summary of notation.

Notation	Definition and description
$\mathbf{X}, \mathbf{Y}_0, \mathbf{M}, \ldots$	Matrices (upper-case)
$\mathbf{U}, \mathbf{V}, \tilde{\mathbf{U}}, \ldots$	Matrix factorization factors (upper-case)
$\mathbf{X}^\top$	Transpose of matrix **X**
X_{ij}	Element at *i*th row, *j*th column of matrix **X**
m, m', n	Dimension of matrix (lower-case)
k	Latent clustering number (lower-case)
ϵ, ξ	Convergence criterions
R_+	Field of non-negative real numbers
$\Omega(\bar{\Omega})$	Observed (missing) data index
$\mathbf{X}_\Omega$	Copying entries from X up to Ω
$\|\| \cdot \|\|_F$	Matrix Frobenius norm
$Tr(\cdot)$	Matrix trace

method (Koren, Bell, & Volinsky, 2009) has been proven successful in the Netflix competition. Multivariate imputation by chained equations (MICE) (van Buuren & Groothuis-Oudshoorn, 2010) is also a popular imputation method which can preserves the observed knowledge during the imputation process. However, these methods are not ideal to handle the missing value imputation, because the original observed information is destroyed in the imputation result. Extended by these methods, our work can reserve the observed information in the imputation result.

2.2. Matrix factorization

Matrix factorization is a widely used method for a large number of data mining problems. It has been widely used and adapted for various purposes such as dimensionality reduction, data clustering, and missing value imputation. The main goal of the MF is to obtain a set of low-rank matrices whose production can approximate the original data with respect to the observed knowledge, as the well-known methods like principal component analysis (PCA) (Jolliffe, 2002) and singular value decomposition (SVD) (Golub & Van Loan, 2012). To be more specific, MF assumes that the partially observed information, i.e. matrix **M**, can be estimated by the product of two low-rank matrices, i.e. matrix **U** and **V**, whose product **UV** shows the minimum Euclidean distance with respect to the observed information in the matrix **M**. The matrices **U** and **V** are treated as factorization factors, meanwhile, the missing value in the matrix **M** is thus estimated in the produce of **UV**.

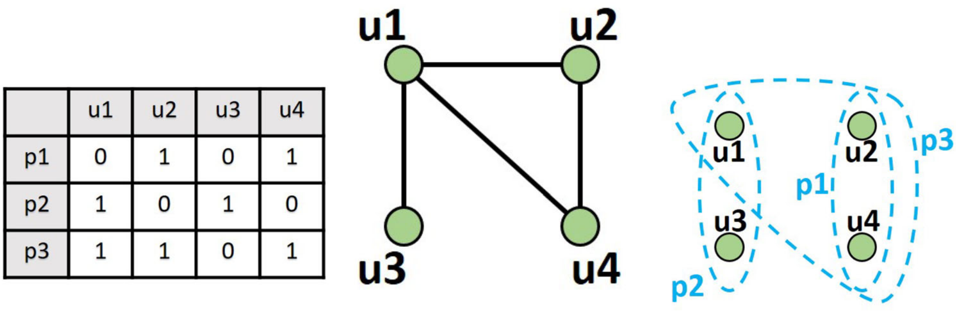
In practice, the non-negative property often exists in many real-world applications, especially for the medical and healthcare domain like medical imaging analysis (Carr, Fright, & Beatson, 1997), gene expression (Gao & Church, 2005), healthcare fraud detection (Zhu, Wang, & Wu, 2011), and medical recommendation system (Zhang, Chen, Huang, Wu, & Li, 2017). These applications naturally require the non-negative property for each entry in the data, however, such non-negative constraint is not satisfied in the MF. To overcome this issue, Lee and Seung (2001) proposes the non-negative matrix factorization (NMF), which has incorporated the non-negative constraint into the MF framework. NMF produces two non-negative low-rank matrices (also known as dual-factors), whose multiplication has the minimum Euclidean distance (defined as the square root of the sum of the absolute squares of the difference between two matrices) regarding the input data. Each of the non-negative low-rank matrices is usually considered as the clustering result for the row-wise and column-wise knowledge of the original data, which reveals the user's emotion and preference in personalized doctor recommendation system (Zhang et al., 2017). The work (Wang & Zhang, 2013) comprehensively reviews the existing NMF methods used in various applications. Meanwhile, a collection of the medical and healthcare data is commonly presented as a patient-by-medical measurement item' matrix, in which the missing entries commonly exist. Several NMF extension methods handle the missing value issue from various perspectives. Xu, Yin, Wen, and Zhang (2012) contributes to recovering the missing data from the partially observed information by taking advantage of MF. Graph regularized non-negative matrix factorization (GNMF) (Cai, He, Han, & Huang, 2011) incorporates the samples' pairwise similarity by introducing the graph regularizer into the traditional NMF to explore insight into the intrinsic geometric structure of the data, which reduce the side effects of the unknown entries. A convex and semi-NMF (Ding, Li, & Jordan, 2010) method expanded the application domain by relaxing the non-negative constraint, and (Wang et al., 2015) incorporates the guidance constraints to align with existing medical knowledge. However, when applying the double orthogonality in dual-factor MF, it is very restrictive and it gives a rather poor matrix low-rank approximation. Thus, Ding, Li, Peng, and Park (2006) proposed the tri-factor factorization method subject to the double orthogonal constraints on both factorization factors, which allows the different cluster number of row and column clustering. Gu, Ding, and Han (2011) proposed to solve the common scale transfer problem by leveraging normalized cut-like constraints. Recent work incorporates the deep learning model with MF for the missing value imputation task (Liu, He, Dubby, & O'Sullivan, 2019).

3. Problem formulation

In this section, we introduce our proposed frameworks for the missing information imputation with heterogeneous auxiliary data, named MI^2-HD and MI^2-HT separately. We first introduce the hypergraph representation for the disease-dedicated social networks and then present how to formulate our goal as an optimization problem. Table 1 summarizes all the frequently used notation in this paper. The boldface uppercase letters denote the matrices (e.g. **X**, **M**). The boldface lowercase letters denote the vectors (e.g. **u**, **v**). $\mathbf{X}_{ij}$ denotes the *i*th row *j*th column entry of matrix **X**, and $\mathbf{u}_i$ denotes the *i*th entry of vector $\boldsymbol{u}$. The regular letters and Greek alphabet are defined as scalars. All vectors are column vectors unless otherwise specified.

3.1. Hypergraph representation of disease-dedicated social networks

The social media networks come naturally in the form of the graph, where the nodes are associated with users, and the edges connecting a pair of users reflect pairwise



	u1	u2	u3	u4
p1	0	1	0	1
p2	1	0	1	0
p3	1	1	0	1

(a) incident matrix. (b) user-user. (c) user-post hypergraph.

Figure 1. Illustration of the hypergraph representation. In sub-figure (a), for example, the user u1 leaves his/her comment in the post p1, as the corresponding value equals to 1 in the incident matrix, otherwise 0. In the sub-figure (b), which indicates the users' grouping information, user u1 and user u2 are connected as both of them have participated in the post p1, while this graph cannot tell us how many users are involved in the same post. The sub-figure (c) is the user-post hypergraph which contains the completed user grouping information of each post on the disease-dedicated social network.

similarity. In the previous studies, the users' pairwise similarity is typically measured as the follower/followee relationship between two users. Two users are usually considered to be similarity when the follower/followee connection is observed. The so-called graph regularized is proposed to quantify the corresponding users' similarity when formulating the objective function in the modeling processing. However, in the context of the disease-dedicated social networks, this pairwise similarity can lead to significant loss of the user grouping information, which reveals purposeful knowledge in the healthcare domain. For example, on the disease-dedicated forum, the graph regularizer is usually used to formulate the users' pairwise similarity, i.e. how these two users similar to each other when concerning their physical/medial measurement like normal blood sugar levels or the level of oral glucose tolerance test. However, usually more than two users reply to the same discussion thread that they are normally suffered from the same type of disease. Multiple users (nodes) are connected by the same post (edge). The patients' pairwise similarity is not able to capture such grouping information. We propose to represent the disease-dedicated social networks by leveraging the hypergraph structure, which can preserve both the pairwise and reveal the grouping knowledge simultaneously.

Different from traditional graphs, in the hypergraph structure, each hyperedge corresponds to one thread (post), and it connects multiple users who have participated in the thread, and thus it is able to effectively preserve the aforementioned grouping information. Case in point, for the online disease-dedicated forum, the first set of users are connected by the thread where they mainly discuss glucose level, while another set of users are connected by the thread where they talk about the insulin pump. These two sets of the users present two different aspects when producing the analysis on the healthcare forum. The hypergraph structure enables the preservation of such grouping information, i.e. the user within the same group (post) are formulated to be close to each other with respect to the hypergraph regularizer, while the user from different group keep relatively far distance from each other. For the users that can be observed in both group (overlapped user), i.e. he/she has attended the discussion in these two threads simultaneously, which can also be reserved in the hypergraph regularizer, but unfortunately eliminated in the traditional graph regularizer. For each post (hyperedge), the hyperedge weight represents the popularity of this post, e.g. how many users' reply or the duration of the discussion. Figure 1 shows a simple example of the hypergraph representation. $User = \{u_1, u_2, u_3, u_4\}$ and $Post = \{p_1, p_2, p_3, p_4, p_5\}$ denote the user set and post set, respectively. The incident matrix in Fig. 1(a) has the entry $(p_i, u_j) = 1$ if user u_j participates in the post p_i; the traditional graph model in Fig. 1(b) shows how the pairs of users are connected when they participate in the same post, while user grouping information for each thread is lost. The traditional graph structure cannot reveal whether the same user left comments under multiple posts, while such kind of grouping knowledge loss is not expected for data mining purposes because the posts with the same user are likely to belong to the same topic, or contain the patients' daily continuous biomarker measurements. The hypergraph in Fig. 1(c) fully describes the user-post grouping relationship when we treat each post as one hyperedge. The connection between each user and the user grouping knowledge for each post is completed illustrated. Thus the high-order relationships among users can be captured by hypergraph structures without loss of any information.

3.2. Proposed MI^2-HD framework

As mentioned before, the traditional pair-wise graph structure does not fully exploit the grouping information existing in the disease-dedicated social networks. Motivated by Cai et al. (2011), which takes the pair-wise similarity into consideration, we further explore the user grouping information by leveraging the hypergraph structure (Zhou, Huang, & Schölkopf, 2007). For the disease-dedicated social network with the number of m users and n posts, let $\mathcal{V}, \mathcal{E}$ denote the use (vertex) set and forum post (hyperedge) set, respectively, then the network can be presented as the hypergraph $G(\mathcal{V}, \mathcal{E}, \mathcal{W})$ with the vertex set $\mathcal{V}$, hyperedge set $\mathcal{E}$, and the hyperedge weight knowledge $\mathcal{W}$. The weighted hypergraph contains the hyperedge weight $w(e) \in \mathcal{W}$ associated with each hyperedge $e \in \mathcal{E}$. For each vertex $v \in \mathcal{V}$, the vertex

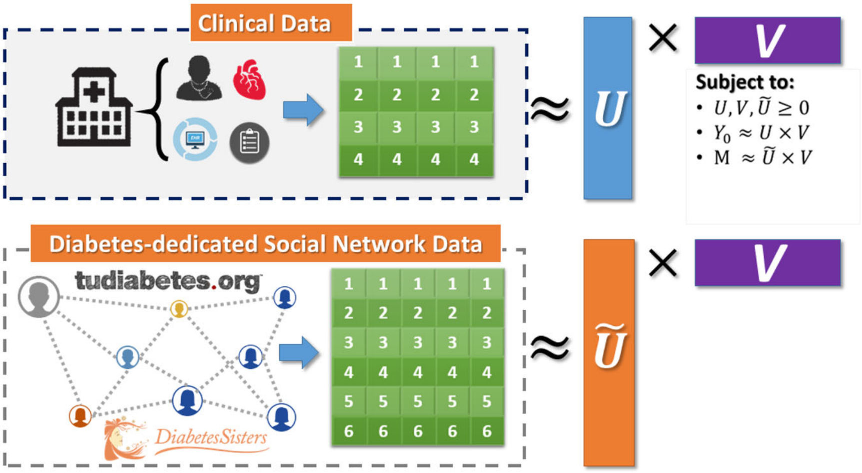


Figure 2. Missing information imputation with auxiliary data. Disease-dedicated social network data provides the patients' self-report information, meanwhile, the high-order user-post relationships can be digged out from it.

degree $d(v)$ is defined as $d(v) = \sum_{\{e \in \mathcal{E} | v \in e\}} w(e)$. The incident matrix $\mathbf{H}$ in the size $|\mathcal{V}| \times |\mathcal{E}|$ indicates whether the user-post connection, that the entry $h(v_i, e_j) = 1$ if $v_i \in e_j$ (user v_i appears in the post e_j), and 0 otherwise. For each hyperedge e, the hyperedge degree $\delta(e)$ is defined as $\delta(e) = \sum_{v \in \mathcal{V}} h(v, e)$, which indicates how many users leave their commons under the post e. The diagonal matrix $\mathbf{D}_v$ in the size $|\mathcal{V}| \times |\mathcal{V}|$ has its diagonal elements equal to the degree of each vertex, and the diagonal matrix $\mathbf{D}_e$ in the size $|\mathcal{E}| \times |\mathcal{E}|$ has its diagonal elements equal to the degree of each hyperedge.

Analogous to the definition of Laplacian matrix in the normal graph (Cai, Mei, Han, & Zhai, 2008), the hypergraph Laplacian matrix $\mathbf{L}^h \in R^{m \times m}$ is defined as $\mathbf{L}^h = \mathbf{D}_v - \mathbf{H}\mathbf{W}_e\mathbf{D}_e^{-1}\mathbf{H}^\top$. By incorporating the hypergraph structure with the MF method, we first formulate our goal as the following optimization problem:

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{V}} \; & ||\mathbf{Y}_0 - \mathbf{U}\mathbf{V}^\top||_F^2 + \lambda Tr(\mathbf{U}^\top \mathbf{L}^h \mathbf{U}) \\ s.t. \quad & \mathbf{U} \geq 0, \mathbf{V} \geq 0 \end{aligned} \tag{1}$$

in which $\mathbf{Y}_0 \in R_+^{m \times n}$ denotes the original clinical data, $\mathbf{U} \in R_+^{m \times k}$ and $\mathbf{V} \in R_+^{n \times k}$ denote the low-rank factorization factor. The original clinical data naturally comes with the missing value. The tradeoff parameter $\lambda \geq 0$ controls the effectiveness of the hypergraph structure. To be more specific, in the Eq. (1) the $Tr(\cdot)$ term incorporates the hypergraph knowledge of the disease-dedicated social network into our model. Following the matrix linear algebra manipulations (Cai et al., 2008; Gao, Tsang, & Chia, 2013), the norm $Tr(\mathbf{U}^\top \mathbf{L}^h \mathbf{U})$ can be rewritten as:

$$\begin{aligned} Tr(\mathbf{U}^\top \mathbf{L}^h \mathbf{U}) &= Tr(\mathbf{U}^\top \mathbf{D}_v \mathbf{U}) - Tr(\mathbf{U}^\top \mathbf{H}\mathbf{W}_e\mathbf{D}_e^{-1}\mathbf{H}^\top \mathbf{U}) \\ &= \sum_{e \in \mathcal{E}} \sum_{(v_i, v_j) \in e} \frac{w(e)}{\delta(e)} ||v_i - v_j||^2 \end{aligned} \tag{2}$$

where the distance between the nodes v_i and v_j within each hyperedge, weighted by the $\frac{w(e)}{\delta(e)}$, is inclined to short. Thus, when minimizing the Eq. (2) as the optimization goal of the Eq. (1), the similarity of the vertices associated with the same hyperedge keeps constant. In other words, considering the practical disease-dedicated forum, the users who share their experience at the same post are expected to be relevant to each other, that this kind of grouping relation is encoded in the Eq. (2) as the similarity among these nodes keeping the same within each hyperedge. Users may discuss different topics in different posts, then the node (user) grouping information is altered regarding the hyperedge (post). For structural convenience, in this paper, we set hyperedge weight to be equal, and the weight effect will be explored in our future.

Secondly, our goal is to impute the missing value in Y_0 (e.g. patient biomarker data) by exploiting the heterogeneous auxiliary information (e.g. clinical trial data, disease-dedicated social network), as shown in Fig. 2, together with strict constraint on the observation information. The goal is motivated by two aspects:

Latent coherence: The latent coherence spreads among the similar users from diverse sample source $\mathbf{M}$ (e.g. diabetic patients), while the samples in $\mathbf{M}$ do not necessarily overlap with the samples in $\mathbf{Y}_0$.

Observation consistency: The dense matrix $\mathbf{Y} = \mathbf{U}\mathbf{V}^\top$ in Eq. (1) is usually treated as the learning result in the previous studies, while the fact is $\mathbf{Y}_\Omega \neq (\mathbf{Y}_0)_\Omega$, where $(\cdot)_\Omega$ index to the observed data in $\mathbf{Y}_0$. The imputation result $\mathbf{Y}$ messes the original observed value in $\mathbf{Y}_0$ due to the inherent divergence of the heterogeneous data. Ideally, we expect the missing entries in $\mathbf{Y}_0$ to be filled up by incorporating the auxiliary information, and meanwhile, keeping $\mathbf{Y}$ consistent with the observed information in $\mathbf{Y}_0$.

Thus, motivated by these ideas, we move forward to extend Eq. (1) by adding the strict constraint on the observed information together with leveraging the heterogeneous auxiliary data. The objective of MI^2-HD is to solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{Y}, \mathbf{U}, \tilde{\mathbf{U}}, \mathbf{V}} \; & ||\mathbf{Y} - \mathbf{U}\mathbf{V}^\top||_F^2 + \alpha||\mathbf{M} - \tilde{\mathbf{U}}\mathbf{V}^\top||_F^2 + \beta Tr(\mathbf{U}^\top \mathbf{L}^h \mathbf{U}) \\ s.t. \; & \mathbf{U} \geq 0, \tilde{\mathbf{U}} \geq 0, \mathbf{V} \geq 0, \mathbf{Y}_\Omega \equiv (\mathbf{Y}_0)_\Omega \end{aligned} \tag{3}$$

$\mathbf{Y} \in R_+^{m \times n}, \mathbf{U} \in R_+^{m \times k}$, $\mathbf{V} \in R_+^{n \times k}, \mathbf{M} \in R_+^{m' \times n}$, $\tilde{\mathbf{U}} \in R_+^{m' \times k}$, and $\mathbf{L}^h \in R^{m \times m}$, where m and m' denote the number of user in original data and auxiliary data, respectively, and n denotes the number of feature. The tradeoff parameters $\alpha, \beta \geq 0$. Matrices $\mathbf{U}$ and $\tilde{\mathbf{U}}$ indicate the row clustering (sample grouping), and matrix $\mathbf{V}$ indicates the column clustering (measurement grouping). The latent coherence among the heterogeneous data $\mathbf{M}$

and original data Y_0 is required by sharing the same measurement matrix $\mathbf{V}$, and simultaneously the missing entries $\mathbf{Y}_{\Omega}$ are updated iterative. $\mathbf{Y}$ and ($\mathbf{U}$, $\mathbf{V}$, $\tilde{\mathbf{U}}$) are updated separately.

3.3. Extension to matrix tri-factorization

Closely related to MI^2-HD, we propose MI^2-HT by leveraging the tri-factor MF model, with non-negativity and orthogonality constraints on each factorization matrix. Compared with the two-factor MF mentioned above, which may provide a relatively weak low-rank approximation (Ding et al., 2006; Wang, Nie, Huang, & Makedon, 2011) introduced one more factorization factor $\mathbf{S}$ into consideration. In this model, the observed data matrix $\mathbf{Y}_0$ is approximated by three factors $\mathbf{U}$, $\mathbf{S}$ and $\mathbf{V}$, that factor S is designed to absorb the different scales of $\mathbf{U}$ and $\mathbf{V}$. In the meanwhile, the auxiliary data $\mathbf{M}$ is also tri-factorized by sharing the same measurement matrix $\mathbf{V}$. The objective function of MI^2-HT is formulated as:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V},\tilde{\mathbf{U}},\mathbf{S},\tilde{\mathbf{S}}} & ||(\mathbf{Y}_0-\mathbf{USV}^\top)_\Omega||_F^2+\alpha||\mathbf{M}-\tilde{\mathbf{U}}\tilde{\mathbf{S}}\mathbf{V}^\top||_F^2+\beta Tr(\mathbf{U}^\top\mathbf{L}^h\mathbf{U}) \\ s.t.\ & \mathbf{U},\tilde{\mathbf{U}},\mathbf{V},\mathbf{S},\tilde{\mathbf{S}}\geq 0,\mathbf{UU}^\top=\mathbf{I},\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top=\mathbf{I},\mathbf{VV}^\top=\mathbf{I} \end{aligned} \tag{4}$$

where $\alpha,\beta\geq 0$, the matrix $\mathbf{M}$ denotes the auxiliary data, collected from the diabetes-dedicated social networks. To avoid ambiguity, the orthogonal constraint on factorization matrices $\mathbf{U}$, $\tilde{\mathbf{U}}$ and $\mathbf{V}$ require only one non-zero entry in each row, which forces each user/biomarker to only belong to a single clustering class.

4. Updating rules

The proposed optimization problem is solved by the joint MF. A set of multiplicative updating rules are proposed to solve the optimization problem.

Algorithm 1 Updating Rules for MI^2-HD

```
Input:
    $\mathbf{Y}_0$, M, $\mathbf{L}^h$, $\mathbf{\Omega}$, $\mathbf{W}_e$, $\mathbf{D}_e$, $\mathbf{D}_v$, H
    initial Y, U, V, and $\tilde{\mathbf{U}}$ as $\mathbf{Y}_0, \mathbf{U}^0, \mathbf{V}^0, \tilde{\mathbf{U}}^0$
    $t = t' = 0, \epsilon = 1, \xi = 1$
Output:
    Y, U, $\tilde{\mathbf{U}}$, V
while $\xi$ >1E-2 do
  $t = 0, \epsilon = 1$
  while $\epsilon$ >1E-2 do
    $U_{ik}^{t+1} \leftarrow U_{ik}^t$ update in Eq. (5)
    $V_{jk}^{t+1} \leftarrow V_{jk}^t$ update in Eq. (5)
    $\tilde{U}_{ij}^{t+1} \leftarrow \tilde{U}_{ij}^t$ update in Eq. (5)
    $ObjValue^{t+1} = O(\mathbf{U}^{t+1}, \mathbf{V}^{t+1}, \tilde{\mathbf{U}}^{t+1})$
    $\epsilon = \frac{ObjValue^{t+1} - ObjValue^t}{ObjValue^t}$
    $t = t + 1$
  end while
  set $\mathbf{Y}_{\Omega}^{t'} = (\mathbf{Y}_0)_{\Omega}$
  set $\xi$ equals to the mean of $(\mathbf{Y}^{t'+1} - \mathbf{Y}^{t'})./\mathbf{Y}^{t'}$
  $t' = t' + 1$
end while
```

4.1. MI^2-HD updating rules

There are two iterative updating steps in MI^2-HD, as shown in Algorithm 1. Since Eq. (3) is convex for the variables $\mathbf{Y}$, $\mathbf{U}$, $\tilde{\mathbf{U}}$, $\mathbf{V}$ separately (See Section 4.4), we propose to update $\mathbf{Y}$ and $\mathbf{U}$, $\tilde{\mathbf{U}}$, $\mathbf{V}$ separately. In the Algorithm 1, the convergence criteria is defined as the average changing rate of the imputation result. When the average changing of the imputation value is less than 1E-2, the updating process is considered as converged.

Fix Y, update U, $\tilde{U}$, V: We first introduce how to update $\mathbf{U}$, $\tilde{\mathbf{U}}$, $\mathbf{V}$ with fixing $\mathbf{Y}$ by minimizing the Eq. (3). The Eq. (3) is then extended into the following form:

$$\begin{aligned} O = & Tr\left[(\mathbf{Y}-\mathbf{UV}^\top)(\mathbf{Y}-\mathbf{UV}^\top)^\top\right] + \beta Tr[\mathbf{U}^\top\mathbf{L}^h\mathbf{U}] \\ & + \alpha Tr\left[(\mathbf{M}-\tilde{\mathbf{U}}\mathbf{V}^\top)(\mathbf{M}^\top-\mathbf{V}\tilde{\mathbf{U}}^\top)\right] \\ = & Tr\left[(\mathbf{Y}_0-\mathbf{UV}^\top)_\Omega(\mathbf{Y}_0^\top-\mathbf{VU}^\top)\right] + \beta Tr(\mathbf{U}^\top\mathbf{L}^h\mathbf{U}) \\ & + \alpha Tr(\mathbf{MM}^\top) \end{aligned}$$

we introduce the Lagrangian function $\mathscr{L}$ and Lagrange multipliers $\mathbf{\Psi}_{ij}$, $\mathbf{Phi}_{ij}$, and $\mathbf{\Gamma}_{ij}$. Each multiplier $\mathbf{\Psi}_{ij}$, $\mathbf{\Phi}_{ij}$, and $\mathbf{\Gamma}_{ij}$ corresponds to the constraints $\mathbf{U}_{ij}\geq 0, \mathbf{V}_{ij}\geq 0$ and $\tilde{\mathbf{U}}_{ij}\geq 0$, respectively. The Lagrange function $\mathscr{L}$ can be written as:

$$\mathscr{L} = O(\mathbf{Y},\mathbf{U},\mathbf{V},\tilde{\mathbf{U}}) + Tr(\mathbf{\Psi}\mathbf{U}^\top) + Tr(\mathbf{\Phi}\mathbf{V}^\top) + Tr(\mathbf{\Gamma}\tilde{\mathbf{U}}^\top)$$

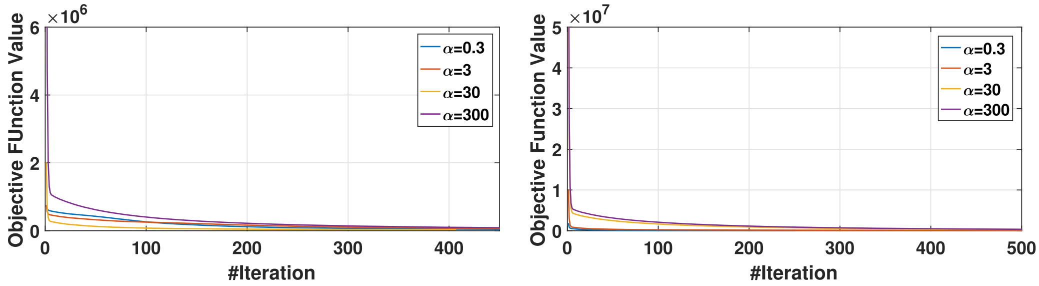
The partial derivatives of $\mathscr{L}$ with respect to U, V and $\tilde{U}$ are:

Algorithm 2 Updating Rules for MI^2-HT

```
Input:
    $\mathbf{Y}_0$, M, $\mathbf{L}^h$
    initial U, V, $\tilde{\mathbf{U}}$, S as $\mathbf{U}^0, \mathbf{V}^0, \tilde{\mathbf{U}}^0, \mathbf{S}^0$
    $t = 0$, $\epsilon = 1$, $Convergence\ Criterion = 10^{-4}$
Output:
    U, V, $\tilde{\mathbf{U}}$, S, $\tilde{\mathbf{S}}$
while $\epsilon > Convergence\ Criterion$ do
  $U_{ik}^{t+1} \leftarrow U_{ik}^t \frac{(\mathbf{YVS}^\top)_{ij}}{(\beta\mathbf{L}^h\mathbf{U}+\beta(\mathbf{L}^h)^\top\mathbf{U})_{ij}}$
  $V_{jk}^{t+1} \leftarrow V_{jk}^t \frac{(\mathbf{Y}^\top\mathbf{US})_{ij}}{(\alpha\mathbf{VS}^\top\mathbf{U}^\top\mathbf{US})_{ij}}$
  $\tilde{U}_{ij}^{t+1} \leftarrow \tilde{U}_{ij}^t \frac{(\mathbf{MV}\tilde{\mathbf{S}}^\top)_{ij}}{(\tilde{\mathbf{U}}\tilde{\mathbf{S}}\mathbf{V}^\top\mathbf{V}\tilde{\mathbf{S}}^\top)_{ij}}$
  $S_{ik}^{t+1} \leftarrow S_{ik}^t \frac{(\mathbf{U}^\top\mathbf{YV})_{ij}}{(\mathbf{V}^\top\mathbf{VS}^\top\mathbf{U}^\top\mathbf{U})_{ij}}$
  $S_{ik}^{\tilde{t+1}} \leftarrow \tilde{S}_{ik}^t \frac{(\tilde{\mathbf{U}}^\top\mathbf{MV})_{ij}}{(\mathbf{V}^\top\mathbf{V}\tilde{\mathbf{S}}^\top\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}})_{ij}}$
  $ObjValue^{t+1} = O(\mathbf{U}^{t+1}, \mathbf{V}^{t+1}, \mathbf{U}^{\tilde{t+1}}, \mathbf{S}^{t+1}, \mathbf{S}^{\tilde{t+1}})$
  $\epsilon = ObjValue^{t+1} - ObjValue^t$
  $t = t + 1$
end while
```

$$\begin{aligned} \frac{\partial\mathscr{L}}{\partial\mathbf{U}} = & -2Tr(\mathbf{YV}) + \frac{\partial Tr(\mathbf{UV}^\top\mathbf{VU}^\top)}{\partial\mathbf{U}} \\ & + \beta\mathbf{LU} + \beta\mathbf{L}^\top\mathbf{U} + \mathbf{\Psi} \\ \frac{\partial\mathscr{L}}{\partial\mathbf{V}} = & -2Tr(\mathbf{YU}) + \frac{\partial Tr(\mathbf{UV}^\top\mathbf{VU}^\top)}{\partial\mathbf{V}} \\ & - \alpha\mathbf{M}^\top\tilde{\mathbf{U}} + \alpha\mathbf{V}\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}} + \mathbf{\Phi} \\ \frac{\partial\mathscr{L}}{\partial\tilde{\mathbf{U}}} = & -2\alpha\mathbf{MV} + 2\alpha\tilde{\mathbf{U}}\mathbf{V}^\top\mathbf{V} + \mathbf{\Gamma} \end{aligned}$$

by setting each partial derivative to 0, based on the Karush–Kuhn–Tucker (KKT) optimality conditions (Boyd & Vandenberghe, 2004) $\mathbf{\Psi}_{ij}\mathbf{U}_{ij} = 0$, $\mathbf{\Phi}_{ij}\mathbf{V}_{ij} = 0$ and $\mathbf{\Gamma}_{ij}\tilde{\mathbf{U}}_{ij} = 0$, we can get:



(a) Convergence of MI^2-HD.

(b) Convergence of MI^2-HT.

Figure 3. Convergence analysis with respect to the tradeoff parameters. The x and y axes denote the iteration number and the objective function value, respectively.

$$(-2\mathbf{YV}+2\mathbf{UV}^\top\mathbf{V}+\beta\mathbf{LU}+\beta\mathbf{L}^\top\mathbf{U})_{ij}\cdot\mathbf{U}_{ij}=0$$
$$(-\mathbf{Y}^\top\mathbf{U}+\mathbf{VU}^\top\mathbf{U}-\alpha\mathbf{M}^\top\tilde{\mathbf{U}}+\alpha\mathbf{V}\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}})_{ij}\cdot\mathbf{V}_{ij}=0$$
$$(-2\alpha\mathbf{MV}+2\alpha\tilde{\mathbf{U}}\mathbf{V}^\top\mathbf{V})_{ij}\cdot\tilde{\mathbf{U}}_{ij}=0$$

Equation (5) leads to the following updating rules for MI^2-HD:

$$\begin{aligned}\mathbf{U}_{ik}&=\mathbf{U}_{ik}\frac{(2\mathbf{YV}+\beta\mathbf{HW}_e\mathbf{D}_e^{-1}\mathbf{H}^\top\mathbf{U})_{ik}}{(2\mathbf{UV}^\top\mathbf{V}+\beta\mathbf{D}_v\mathbf{U}+\beta\mathbf{D}_v^\top\mathbf{U})_{ik}}\\ \mathbf{V}_{kj}&=\mathbf{V}_{kj}\frac{(2\mathbf{Y}^\top\mathbf{U}+\alpha\mathbf{M}^\top\tilde{\mathbf{U}})_{kj}}{2(\mathbf{VU}^\top\mathbf{U}+\alpha\mathbf{V}\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}})_{kj}}\\ \tilde{\mathbf{U}}_{ij}&=\tilde{\mathbf{U}}_{ij}\frac{(\mathbf{MV})_{ij}}{(\tilde{\mathbf{U}}\mathbf{V}^\top\mathbf{V})_{ij}}\end{aligned}\tag{5}$$

Fix U, $\tilde{\mathbf{U}}$, V, update Y: After the convergence of $\mathbf{U}$, $\tilde{\mathbf{U}}$, $\mathbf{V}$, we set $\mathbf{Y}_{\boldsymbol{\Omega}}=(\mathbf{Y}_0)_{\boldsymbol{\Omega}}$ to restore the observed information. Then repeating update $\mathbf{U}$, $\tilde{\mathbf{U}}$, $\mathbf{V}$ until $\mathbf{Y}$ converge.

4.2. MI^2-HT updating rule

Derivation of Eq. (4) follows the same procedure as the derivation of Eq. (3). Omitted for brevity, the updating rules for MI^2-HT are directly given in Algorithm 2.

4.3. Time complexity

The main computational cost is due to matrix multiplication. Therefore, omitted for space, the time complexity for MI^2-HD is $O(m^2n^2p^2k)$, and MI^2-HT is $O(m^2n^2k)$, where m and n denotes the number of user and feature, respectively. The scalar p denotes the number of the post on the disease-dedicated forum. The number k is the latent factorization dimension. The time complexity is empirically verified in the experiment section.

4.4. Convergence analysis

The Algorithm 1 is not jointly convex for all the variables $\mathbf{Y}$, $\mathbf{U}$, $\tilde{\mathbf{U}}$, $\mathbf{V}$, but convex in each of them separately. As shown in Eq. (3), when $\mathbf{Y}$ is fixed, the proof regarding the convexity of Eq. (3) with respect to variables $\mathbf{U}$, $\tilde{\mathbf{U}}$, $\mathbf{V}$ is analogous to (Cai et al., 2008); when $\mathbf{U}$, $\tilde{\mathbf{U}}$, $\mathbf{V}$ are fixed, the optimization problem is equivalent to $\min_{\mathbf{Y}}||\mathbf{Y}-\mathbf{C}||_F^2$, s.t. $\mathbf{Y}_{\boldsymbol{\Omega}}=(\mathbf{Y}_0)_{\boldsymbol{\Omega}}$, with respect to $\mathbf{Y}$ only. $\mathbf{C}$ is given as constant. To be more specific, the equality constraint can be rewritten as $||\boldsymbol{\Lambda}\mathbf{Y}^C-\mathbf{C}||_F^2$, where $\mathbf{Y}^C$ denotes the column-wise concatenation of $\mathbf{Y}$, and $\boldsymbol{\Lambda}$ is a constant diagonal matrix with the diagonal elements equal to the column-wise concatenation of $\boldsymbol{\Omega}$. Thus, the local optima are feasible when Algorithm 1 is proved to be convex with respect to variables $\mathbf{Y}$, $\mathbf{U}$, $\tilde{\mathbf{U}}$, and $\mathbf{V}$ individually. The same proof procedure can be easily adapted to proof that Algorithm 2 also achieves the local optimal solution.

5. Experimental results and discussion

In this section, we evaluate the performance of our missing value imputation method and demonstrate its effectiveness by leveraging the hypergraph structure together with heterogeneous auxiliary data. The computation environment contains Intel(R) CORE 3.50GHz CPU, 32GB RAM in MATLAB R2017b.

The experiments are conducted on the one synthetic data and two real-world data sets. For each data set, the proposed method is compared with four baseline methods: *NMF* (Lee & Seung, 1999), *GNMF* (Cai et al., 2011), *MF-NMF* (Xu et al., 2012) and the traditional regularized expectation-maximization based method RegEm (Schneider, 2001). We use F-norm in *NMF*, measure the pairwise similarity based on Euclidean distance for *GNMF*, and follow the parameter setting in (Xu et al., 2012) for *MF-NMF*.

5.1. Synthetic data

As mentioned in the reference (Hofmann, 2003), the Gaussian distribution is adopted to estimate the users' rating for the item when studying the user preferences, in which each community can be identified by a Gaussian distribution generated from the normalized user ratings. In our case, when we generate the synthetic data, we assume that each user can also be identified by a Gaussian distribution, which is generated according to the user's attendance to each post (topic). For example, a certain group of the users has the frequent participant in the post aiming at discussing the daily diabetes measurements, e.g. blood sugar level and glucose level, while seldom attend the discussion about diabetes

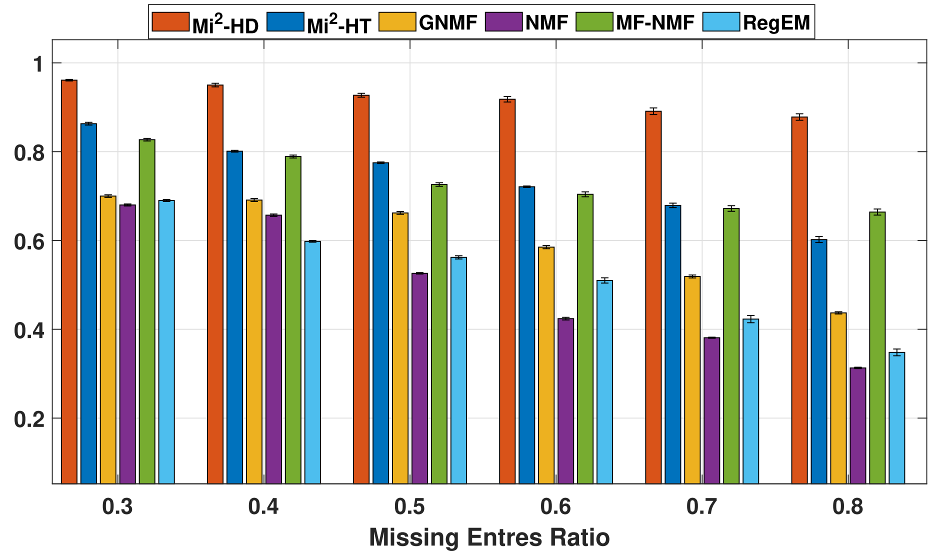


Figure 4. Comparison analysis on synthetic data. The first two bar of each bar-group represent the imputation accuracy of MI^2-HD and MI^2-HT.

pathogenesis (which may attract another set of certain users like the physician, medical practitioner or doctor). Thus, the different group of the user can be distinguished by the Gaussian distribution, which is used to formulate their online activities pattern.

Base on this observation, the factorization factors $\mathbf{U}$, $\mathbf{V}$, and $\tilde{\mathbf{U}}$ are generated based on the multivariate Gaussian distribution. Each of them contains 200, 450 and 150 examples, respectively. The convergence performance is presented in Fig. 3. The tradeoff parameter is selected in grid $[0.3, 3, 30, 300]$ to balance the effectiveness of each regularization term and avoid single term monopolizing the objective function. The multiplicative updating rule shows the robust performance with respect to various parameters settings.

We also consider the scenario when the data sparsity spreads over different sparse ratio. The data sparsity ratio alters the imputation accuracy of our methods. In Fig. 4, the imputation value accuracy is compared with three other algorithms on synthetic data. The x-axis represents the missing ratio, and the y-axis shows the accuracy of the imputation value. It can be observed that the imputation performance is much more stable when the ratio of missing fraction getting increasing. We take cosine similarity Nguyen and Bai (2010) to measure the imputation accuracy between the imputation result and original value. To be more precise, cosine similarity is a measurement that measures the similarity between two non-zero vectors by calculating their cosine value of the angle in their inner product space. Each result is the average over 30-run results. For each running, $\mathbf{U}$, $\mathbf{V}$, and $\tilde{\mathbf{U}}$ are randomly initialized from multivariate Gaussian distribution. Compared with the other three methods, whose accuracies decline quickly along with the increasing ratio of missing entries, MI^2-HD algorithm decreases relatively slow and shows the highest imputation accuracy.

We evaluated the efficiency of the proposed techniques. The results are based on the synthetic data set with varying data volume m and n, respectively. As shown in Fig. 5(a,b), the running time of MI^2-HD and MI^2-HT are *quadratic* with respect to parameter m, MI^2-HD is *quadratic* with respect to parameter n, and MI^2-HT is *quartic* with respect to parameter n, which is consistent with the time complexity analysis in Section 4.

5.2. Real data

In this subsection, we present the experimental results on two real data sets, including one collected from a diabetes-dedicated social network.

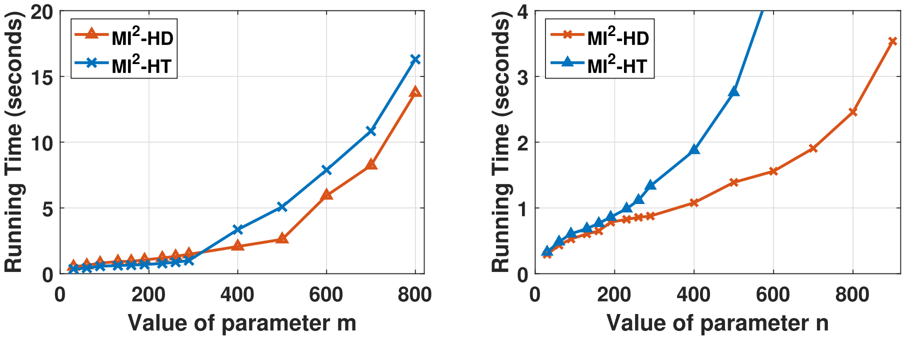
5.2.1. Data set description

There are two real-world data sets used in our experiments:

TuDiabetes data set: The TuDiabetes online forum consists of a community of people touched by diabetes and the disease-specific discussion about their diabetes condition. The set of discussions usually include type I diabetes, type II diabetes, gestational diabetes, diet, exercise, etc. As the screenshot of the TuDiabetes forum shown in Fig. 6, the users tend to form the same groups with interest in a certain topic. In general, the Tudiabetes data set is a collection of 21,286 discussion posts with 294,272 users. The features for each user consist of the TF-IDF (Salton & Yang, 1973) feature of his/her posts after the pre-processing steps (verb tense uniform, stop word removal).

OneID: The OneID data set (Zhongqi, 2015) contains the encrypted user online shopping activities, including the device-cookie pair, searching keywords, auction ID, shop ID and so on. The users' feature is extracted by the Geohash method (Geohashes, 2008) from the raw encrypted information that each feature is converted into a vector of the same length. For detailed information, readers are recommended to see the reference.

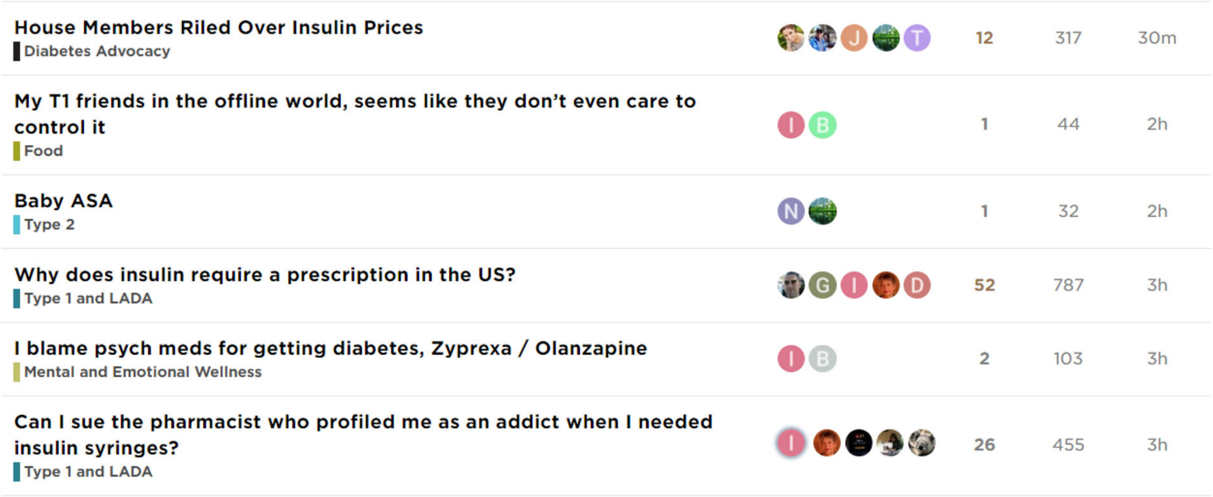
To be more specific, the posts with less than two users are screened out for the purpose of the high-quality hypergraph construction. Each data set is randomly partitioned into two subsets consisting of 80% and 20% of the whole data set, respectively. The 80% subset is treated as Y_0, and



(a) Running time v.s. m.

(b) Running time v.s. n.

Figure 5. Experimental analysis: (a) imputation improvement by leveraging hypergraph structure; (b) imputation improvement by utilizing heterogeneous data; (c, d) running time analysis.



House Members Riled Over Insulin Prices Diabetes Advocacy	J T	12	317	30m
My T1 friends in the offline world, seems like they don't even care to control it Food	I B	1	44	2h
Baby ASA Type 2	N	1	32	2h
Why does insulin require a prescription in the US? Type 1 and LADA	G I D	52	787	3h
I blame psych meds for getting diabetes, Zyprexa / Olanzapine Mental and Emotional Wellness	I B	2	103	3h
Can I sue the pharmacist who profiled me as an addict when I needed insulin syringes? Type 1 and LADA	I	26	455	3h

Figure 6. The screen shot of the TuDiabetes forum.

the other subset, with added noise, is used to simulate the heterogeneous auxiliary information source M.

5.2.2. Experimental results

In this subsection, we present the experimental results on the real work data sets and the effectiveness of our methods.

To evaluate the effectiveness of the proposed algorithm on missing value imputation with the heterogeneous auxiliary information, we use a range of missing ratios (*mRatio*) to partition Y_0 into the observed portion and the missing portion. The missing entries are randomly selected based on the value of *mRatio* in the grid $[0.3, 0.4, 0.5, 0.6, 0.7, 0.8]$, e.g. $mRatio = 0.4$ means 40% entries in Y_0 are manually removed as missing, and replaced with value 0. The removed portion is used as ground truth to evaluate the imputation accuracy.

The 30-run average results on both the TuDiabetes data set and the OneID data set are shown in Fig. 7. The first two bars of each bar-group represent the imputation accuracy for our MI^2-HD and MI^2-HT. Overall, with the increasing of missing value fraction, our method shows stable high accuracy with the help of hypergraph structure and the heterogeneous auxiliary information. To be explicit, the experiment results verify the two main advantages of our method:

1. As shown in Fig. 8(a), by leveraging the hypergraph structure, the proposed MI^2-HD can improve the missing value imputation performance when compared with the traditional graph-based methods *GNMF*. Compared with the model MI^2-HT, the model MI^2-HD shows higher missing value imputation accuracy on both data sets. The reason is that in model MI^2-HT, the strict orthogonality constraints have been adopted on the factorization factors, i.e. $\mathbf{U}$, $\tilde{\mathbf{U}}$, and $\mathbf{V}$. The model MI^2-HT benefits from the mathematical property of the orthogonality constraint, which reduces the computational complexity dramatically. However, such an orthogonality constraint presents the one-to-one mapping relationship among the users and posts, which ignoring the user-grouping knowledge of the disease-dedicated social networks, even though we have addressed the user-grouping knowledge in the Eq. (4) by leveraging the hypergraph structure. The proper constraints give rise to better imputation accuracy, which we will attach great importance in our future works.
2. As shown in Fig. 8(b), the superiority of our methods is increasing along with the data sparsity growing up. The imputation accuracy and robustness are benefited from utilizing the heterogeneous data by sharing the

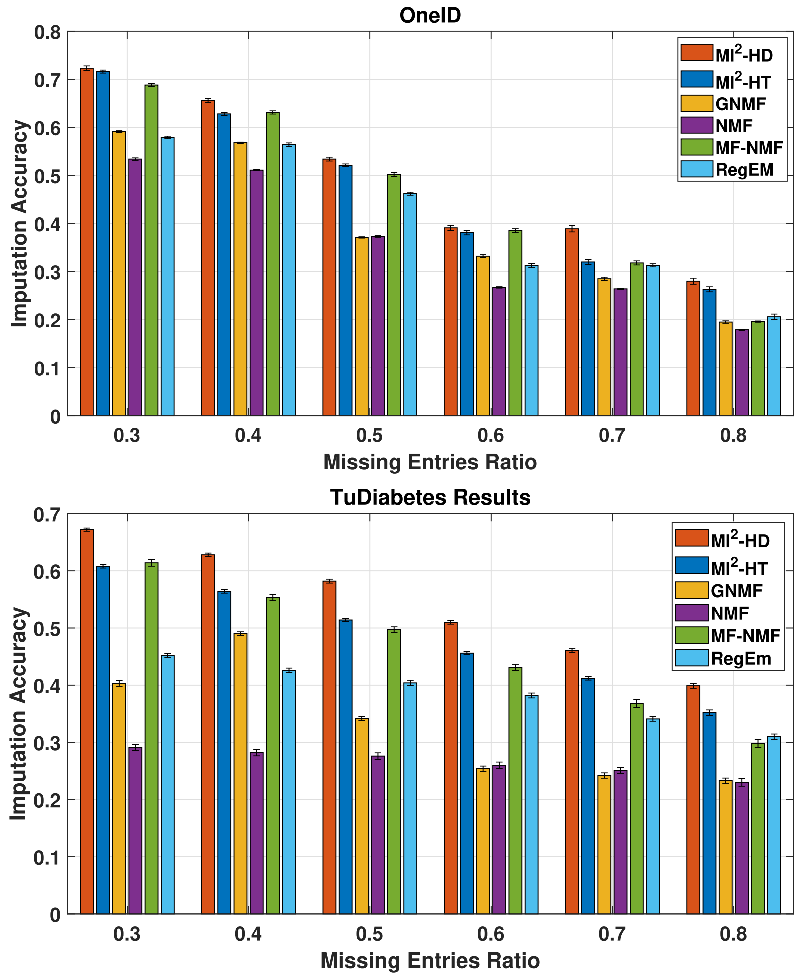
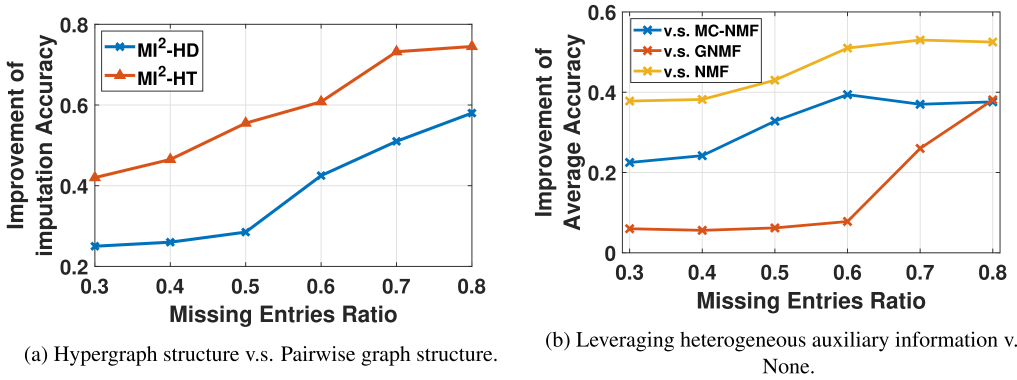


Figure 7. Upper: experimental results on the OneID dataset; lower: experimental results on the TuDiabetes dataset. Each numerical value is averaged over 30-run repeated test, then the 30-run variance is shown in the error bar.



(a) Hypergraph structure v.s. Pairwise graph structure.

(b) Leveraging heterogeneous auxiliary information v.s. None.

Figure 8. Experimental analysis: (a) imputation improvement by leveraging hypergraph structure. (b) Imputation improvement by utilizing heterogeneous data.

measurement matrix V between the original and heterogeneous information.

5.3. Discussion

One experimental setting that needs to be addressed is that each hyperedge is considered of equal importance with the others, as all the hyperedges are given equal weight. In practice, that means each post is considered of the same importance, however, regardless of the fact that the posts' popularity is altered regarding the users and topics. For example, the post titled '*Insulin not bringing my blood sugar DOWN?*', which has been viewed over 56.3k times, is more popular than the post about '*Can't log into Dexcom Clarity*', which has been viewed about 257 times. Conspicuously, users who attended the former discussion are more likely to provide diabetes-related information, rather than the users who discussed the online system login issue. As each hyperedge represents one post, the equal weight setting may cause bias regarding this observation. There are two ways we have thinking about to bring in the posts' popularity and distinguish the user grouping information:

- The most native way is to weight each post (hyperedge) according to its activities, e.g. the numbers of reply, or how many times it has been viewed. The more it has been discussed/viewed represents the potential importance of this topic among the diabetes patients, that the users (nodes) within the corresponding post (hyperedge) should be high-valued relatively.
- A more reasonable way to weight the post (hyperedge) is to conduct the linguistic analysis for the topic of each post. The previous native way is applicable in the most cases, however, both the posts '*Convenia: A Dangerous Veterinary Drug: Please don't ever use this drug for your cats and dogs!*' and '*High blood sugar causing chest pain?*' have been viewed around 29k times and show the similar popularity on the disease-dedicated forum. The basic natural language processing methods like TF-IDF, Latent Dirichlet Allocation (LDA) would be helpful to make a distinguished judgment for the topic of each post and set the proper weight for each post.

To be more specific, the equal weight of the hyperedge does not affect the grouping information. The grouping information is presented at the node (user) level, which means the nodes (users) associated with the same hyperedge (post) are expected to be relevant to each.

6. Conclusion

In this paper, we propose a novel framework for missing value imputation with heterogeneous auxiliary information named MI^2-HD. It is based on the hypergraph representation of disease-dedicated social networks and leverages additional information such as users' social relationships and clinical data to improve the accuracy of missing value imputation. Furthermore, we propose iterative algorithms to solve the resulting optimization problems and analyze their performance from multiple perspectives. Experimental results show that the proposed techniques are able to outperform state-of-the-art approaches on both synthetic and real data sets collected from diabetes-dedicated social networks.

Acknowledgement

This work is supported by National Science Foundation under Grant No. IIS-1947203 and Grant No. IIS-1813464, the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-02-00 and Ordering Agreement Number HSHQDC-16-A-B0001, and an IBM Faculty Award. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Cai, D., He, X., Han, J., & Huang, T. S. (2011). Graph regularized non-negative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(8), 1548–1560. doi:10.1109/TPAMI.2010.231

Cai, D., Mei, Q., Han, J., & Zhai, C. (2008). Modeling hidden topics on document manifold. *Proceeding of the 17th ACM conference on Information and Knowledge Management (CIKM'08)* (pp. 911–920). doi:10.1145/1458082.1458202

Carr, J. C., Fright, W. R., & Beatson, R. K. (1997). Surface interpolation with radial basis functions for medical imaging. *IEEE Transactions on Medical Imaging*, *16*(1), 96–107. doi:10.1109/42.552059

Ding, C. H., Li, T., & Jordan, M. I. (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 45–55. doi:10.1109/TPAMI.2008.277

Ding, C., Li, T., Peng, W., & Park, H. (2006). Orthogonal nonnegative matrix t-factorizations for clustering. *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 126–135), ACM. doi:10.1145/1150402.1150420

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10), 1087–1091. doi:10.1016/j.jclinepi.2006.01.014

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Gao, Y., & Church, G. (2005). Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, *21*(21), 3970–3975. doi:10.1093/bioinformatics/bti653

Gao, S., Tsang, I. W.-H., & Chia, L.-T. (2013). Laplacian sparse coding, hypergraph Laplacian sparse coding, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 92–104. doi:10.1109/TPAMI.2012.63

Golub, G. H., & Van Loan, C. F. (2012). *Matrix computations* (Vol. 3). Baltimore: JHU Press.

Greenland, S., & Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, *142*(12), 1255–1264. doi:10.1093/oxfordjournals.aje.a117592

Gu, Q., Ding, C., & Han, J. (2011). On trivial solution and scale transfer problems in graph regularized NMF. *IJCAI Proceedings-International Joint Conference on Artificial Intelligence* (Vol. 22, p. 1288).

Hofmann, T. (2003). Collaborative filtering via Gaussian probabilistic latent semantic analysis. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 259–266), ACM. doi:10.1145/860435.860483

Janssen, K. J., Donders, A. R. T., Harrell Jr F. E., Vergouwe, Y., Chen, Q., Grobbee, D. E., & Moons, K. G. (2010). Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology*, *63*(7), 721–727. doi:10.1016/j.jclinepi.2009.12.008

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer Magazine.*, *42*(8), 30–37. doi:10.1109/MC.2009.263

Lange, T., & Buhmann, J. M. (2006). Fusion of similarity data in clustering. *Advances in Neural Information Processing Systems*, *18*, 723–730.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. doi:10.1038/44565

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, *9*, 556–5620.

Liu, X., He, J., Duddy, S., & O'Sullivan, L. (2019). Convolution-consistent collective matrix completion. *Proceedings of the 28th ACM* *international conference on information and knowledge management* (pp. 2209–2212), ACM. doi:10.1145/3357384.3358111

Luengo, J., García, S., & Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, *32*(1), 77–108. doi:10.1007/s10115-011-0424-2

Musil, C. M., Warner, C. B., Yobas, P. K., & Jones, S. L. (2002). A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research*, *24*(7), 815–829. doi:10.1177/019394502762477004

Naming Geohashes. (2008). Geohash. http://geohash.org/site/tips.html

Nguyen, H. V., & Bai, L. (2010). Cosine similarity metric learning for face verification. *Asian conference on computer vision* (pp. 709–720), Springer.

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, *9*, 157–166. doi:10.2147/CLEP.S129785

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–96.

Salton, G., & Yang, C.-S. (1973). On the specification of term values in automatic indexing. *Journal of Documentation*, *29*(4), 351–372. doi:10.1108/eb026562

Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, *14*(5), 853–871. doi:10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2

van Buuren, S., & Groothuis-Oudshoorn, K. (2010). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–68. doi:10.18637/jss.v045.i03

Wang, Y., Chen, R., Ghosh, J., Denny, J. C., Kho, A., Chen, Y., … Sun, J. (2015). Rubik: Knowledge guided tensor factorization and completion for health data analytic. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1265–1274), ACM. doi:10.1145/2783258.2783395

Wang, H., Nie, F., Huang, H., & Makedon, F. (2011). Fast nonnegative matrix tri-factorization for large-scale data co-clustering. *IJCAI Proceedings-International Joint Conference on Artificial Intelligence* (Vol. 22, p. 1553).

Wang, Y.-X., & Zhang, Y.-J. (2013). Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, *25*(6), 1336–1353. doi:10.1109/TKDE.2012.51

Wells, B. J., Chagin, K. M., Nowacki, A. S., & Kattan, M. W. (2013). Strategies for handling missing data in electronic health record derived data. *EGEMS (Generating Evidence & Methods to Improve Patient Outcomes)*, *1*(3), 7. doi:10.13063/2327-9214.1035

Xu, Y., Yin, W., Wen, Z., & Zhang, Y. (2012). An alternating direction algorithm for matrix completion with nonnegative factors. *Frontiers of Mathematics in China*, *7*(2), 365–384. doi:10.1007/s11464-012-0194-5

Zhang, Y., Chen, M., Huang, D., Wu, D., & Li, Y. (2017). idoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. *Future Generation Computer Systems*, *66*, 30–35. doi:10.1016/j.future.2015.12.001

Zhongqi, L. (2015). Rec-Tmall. //tianchi.aliyun.com/datalab/dataSet.htm?id = 2

Zhou, D., Huang, J., & Schölkopf, B. (2007). Learning with hypergraphs: Clustering, classification, and embedding. *Advances in Neural Information Processing Systems*, *19*, 1601–1608.

Zhu, S., Wang, Y., & Wu, Y. (2011). Health care fraud detection using nonnegative matrix factorization. 2011 6th International Conference on Computer Science & Education (ICCSE) (pp. 499–503), IEEE.