Joint Learning of Phenotypes and Diagnosis-Medication Correspondence via Hidden Interaction Tensor Factorization

Kejing Yin[1], **William K. Cheung**[1], **Yang Liu**[1], **Benjamin C. M. Fung**[2] and **Jonathan Poon**[3]
[1] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[2] School of Information Studies, McGill University, Montreal, Canada
[3] Hong Kong Hospital Authority, Hong Kong SAR, China
{cskjyin, william, csygliu}@comp.hkbu.edu.hk, ben.fung@mcgill.ca, jonathan@ha.org.hk

Abstract

Non-negative tensor factorization has been shown effective for discovering phenotypes from the EHR data with minimal human supervision. In most cases, an interaction tensor of the elements in the EHR (e.g., diagnoses and medications) has to be first established before the factorization can be applied. Such correspondence information however is often missing. While different heuristics can be used to estimate the missing correspondence, any errors introduced will in turn cause inaccuracy for the subsequent phenotype discovery task. This is especially true for patients with multiple diseases diagnosed (e.g., under critical care). To alleviate this limitation, we propose the hidden interaction tensor factorization (HITF) where the diagnosis-medication correspondence and the underlying phenotypes are inferred simultaneously. We formulate it under a Poisson non-negative tensor factorization framework and learn the HITF model via maximum likelihood estimation. For performance evaluation, we applied HITF to the MIMIC III dataset. Our empirical results show that both the phenotypes and the correspondence inferred are clinically meaningful. In addition, the inferred HITF model outperforms a number of state-of-the-art methods for mortality prediction.

1 Introduction

Electronic health records (EHR) contain rich clinical data about patients, including diagnoses, prescription orders, laboratory test results, etc. Strategic use of them can accelerate clinical research, and improve healthcare quality [Jensen *et al.*, 2012]. However, the raw data of EHR have a lot of missing information, and are frequently inaccurate, highly complex, and possibly biased [Hripcsak and Albers, 2013]. This hinders the reliability of taking the data-driven paradigm for clinical research. Therefore, the raw EHR data are often mapped to some clinically meaningful and interpretable concepts, which are typically referred to as *phenotypes*. Traditional approaches for phenotyping are based on supervised learning, in which the medical experts specify some target diseases, assign the class label for patient samples, and manually define the features [Lasko *et al.*, 2013]. This approach is known to be time-consuming and labor-intensive [Hripcsak and Albers, 2013]. Various machine learning methods have been proposed to automatically discover multiple phenotypes from the EHR data with minimal human supervision [Yu *et al.*, 2015; Ravì *et al.*, 2017; Wang *et al.*, 2015; Kim *et al.*, 2017]. Among them, non-negative tensor factorization (NTF) has been shown to be effective, especially for the structured EHR data, with the capability of preserving and modeling the interaction structures, which typically cannot be achieved by non-NTF based methods, such as matrix decomposition [Ho *et al.*, 2014b; Wang *et al.*, 2015; Yang *et al.*, 2017; Kim *et al.*, 2017]. For instance, the *patient-diagnoses-medication* interaction can be modeled using a third-order tensor $\mathcal{X}$ where the entry $x_{ijk} = c$ denotes that medication k is prescribed c times to patient i in response to diagnosis j. Interpretable latent patterns as representations of the underlying phenotypes can be automatically discovered via non-negative tensor factorization.

To apply tensor factorization, we need to first define the tensor based on the interaction information which, however, is often not available in the EHR data. It is typical that only a list of diagnoses and a list of medications are recorded per clinical visit, with their correspondence totally missing. Existing methods take the "equal-correspondence" strategy and construct the tensor by assuming all those diagnoses and medications per visit to be equally corresponding to each other. Some assume the *diagnosis-medication* correspondence to be binary while some assign the recorded counts of the medications to all the recorded diagnoses. Fig. 1(a) illustrates the correspondence based on such assumption, where Metoprolol is assumed to correspond equally to both hypertension and pneumonitis since they co-occurred in the records. However, it is known that Metoprolol is typically used to treat hypertension but not pneumonitis in clinical practice. HITF, as depicted in Fig. 1(b), can correctly infer that Metoprolol is corresponding to hypertension only, which is consistent with the medical knowledge and practice. It shows that the "equal-correspondence" strategy will inevitably cause errors. For datasets like MIMIC-III which is a critcal care database,

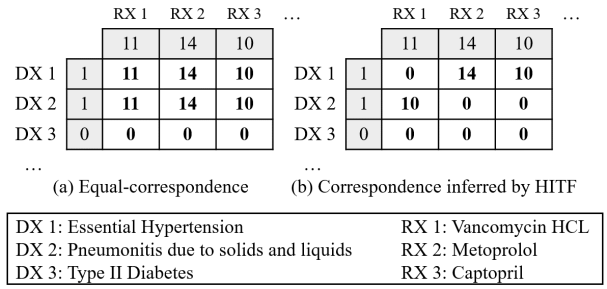

Figure 1: Illustration of diagnosis-medication correspondence: Each row denotes a disease ($DX1/DX2/DX3$) and the "1/0" value next to it indicates if the disease is diagnosed or not. Each column denotes a medication ($RX1/RX2/RX3$) and the number underneath each medication denotes the amount of prescribed medications. (a) Adopting equal-correspondence strategy. (b) Correspondence inferred by the proposed HITF model, which is more reasonable.

multiple diagnoses and medications are often recorded per clinical visit. The resulting inaccuracy will become significantly high.

In this paper, we propose a novel tensor factorization method called *Hidden Interaction Tensor Factorization* (HITF) where the aforementioned correspondence information is estimated together with the tensor factorization. In particular, given only a *patient-by-diagnosis* binary matrix $\mathbf{D}'$ and a *patient-by-medication* counting value matrix $\mathbf{M}$, we take the Poisson CP tensor factorization of the hidden interaction tensor representing the *patient-diagnosis-medication* interaction to discover the underlying phenotypes. We evaluate HITF on the MIMIC-III dataset [Johnson *et al.*, 2016]. The empirical results obtained show that the proposed HITF model can achieve better performance, both quantitatively and qualitatively when compared with the state-of-the-art non-negative tensor factorization based phenotyping models. To the best of our knowledge, this is the first model using tensor factorization with all the unobserved entries of the interaction tensor inferred from the EHR data.

2 Related Work

Applying tensor factorization to the healthcare domain has been intensively studied in the past decade for applications like computational phenotyping [Luo *et al.*, 2016]. In [Ho *et al.*, 2014a], a non-negative tensor factorization method was applied to discover multiple phenotypes from the EHR data. This method was then extended by adding a bias component to capture the baseline characteristics among the overall population [Ho *et al.*, 2014b], and by imposing domain (medical) knowledge into a guidance matrix [Wang *et al.*, 2015]. More recently, some information existing in the EHR data set like in-hospital mortality or medical cost was leveraged to make the phenotypes more discriminative [Yang *et al.*, 2017]. Also, some clustering structure can be added to further ensure the discovered phenotypes being more distinct from each other [Kim *et al.*, 2017].

Regarding the aforementioned missing data problem, it is in fact common for the input tensor data to be partially missing, triggering a large volume of studies on tensor completion via tensor factorization [Liu *et al.*, 2014]. Some recent study also considered the situation that partial indices of the tensor are missing [Yamaguchi and Hayashi, 2017], and yet the existence of at least part of the tensor entries, *i.e.,* interactions being observed, are typically assumed. Different from the existing tensor completion and index inference methods, we propose to jointly infer the interaction together with the latent factors. The work most similar to ours is [Gunasekar *et al.*, 2016], which performs non-negative matrix factorization over several matrices with one shared dimension. Instead of decomposing the matrices separately, we model the interactions by a hidden interaction tensor explicitly and make use of the information across all the dimensions.

3 Notations and Preliminaries

In this paper, we denote the interaction tensor and the reconstructed tensor by $\mathcal{X}$ and $\hat{\mathcal{X}}$, and the (i,j,k)-th entry by x_{ijk} and $\hat{x}_{ijk}$ respectively. The factor matrix associated with the n-th dimension is denoted by $\mathbf{U}^{(n)}$, with its r-th column vector by $\mathbf{u}_r^{(n)}$ and its (i,j)-th entry by $u_{ij}^{(n)}$ respectively. The observed matrices, i.e., *patient-by-medication* matrix and binarized *patient-by-diagnosis* matrix are denoted by $\mathbf{M}$ and $\mathbf{D}'$ respectively. The number of patients, diagnoses and medications are N_p, N_d and N_m respectively.

CP Decomposition. The CP decomposition [Kolda and Bader, 2009] approximates the input tensor with the sum of component rank-one tensors, where each of which can be interpreted as one latent factor. For example the CP decomposition of a third-order tensor $\mathcal{X}$ is defined as follows:

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{u}_r^{(1)} \circ \mathbf{u}_r^{(2)} \circ \mathbf{u}_r^{(3)} = [\![\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}]\!], \quad (1)$$

where R is the number of rank-one tensors.

Slices. Two-dimensional slices are sections of a tensor and are obtained by fixing all but two indices. For a third-order tensor $\mathcal{X}$, together with its CP decomposition defined as aforementioned, the slice with the index of the second dimension fixed at j can be written as [Kolda and Bader, 2009]:

$$\mathbf{X}_{:j:} = \mathbf{U}^{(1)} \operatorname{diag}(\mathbf{u}_{j:}^{(2)}) {\mathbf{U}^{(3)}}^T, \quad (2)$$

where $\operatorname{diag}(\cdot)$ is the diagonal operator that takes a vector as input and gives a diagonal matrix with the elements of the input vector on the main diagonal as output.

Accumulation. We define the accumulation of a tensor as a matrix obtained by summing all slices of the tensor along the same dimensions. For a third-order tensor $\mathcal{X}$, the accumulation along the second dimension is:

$$\mathbf{M} = \sum_{j=1}^{J} \mathbf{X}_{:j:} = \mathbf{U}^{(1)} \operatorname{diag}(\mathbf{1}^T \mathbf{U}^{(2)}) {\mathbf{U}^{(3)}}^T, \quad (3)$$

where $\mathbf{1}$ is the vector of all ones.

4 Proposed Model

4.1 Formulation

Given a *patient-by-medication* matrix $\mathbf{M}$ and a binary *patient-by-diagnosis* matrix $\mathbf{D}'$, our goal is to jointly infer the diagnosis-medication interactions and the latent phenotypes.

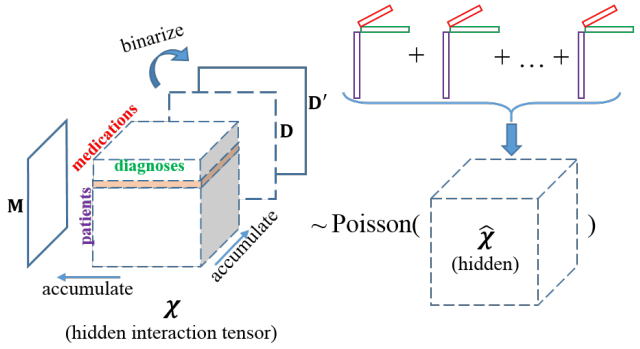

Figure 2: Framework of the proposed model: Only the accumulation along medication mode and diagnoses mode are observed. The hidden interaction tensor are assumed to be drawn from a Poisson distribution where the mean is reconstructed from the tensor factors.

We formulate the problem as a Poisson non-negative tensor factorization (Poisson NTF) problem, where the interactions among patients, diagnoses and medications are explicitly modeled with a hidden *patient-diagnosis-medication* tensor $\mathcal{X}$, with the (i,j,k)-th entry x_{ijk} denoting the amount of medication k being prescribed to patient i in response to the diagnosis j. The framework of our proposed model is illustrated in Fig. 2. In Poisson non-negative tensor factorization, the entries of the input tensor $\mathcal{X}$ are assumed to be drawn from a Poisson distribution, where the mean is the reconstructed tensor $\hat{\mathcal{X}}$ [Chi and Kolda, 2012], i.e.,

$$x_{ijk} \sim \text{Poisson}(\hat{x}_{ijk}). \tag{4}$$

The standard Poisson NTF model solves for the CP factor matrices by maximizing the likelihood of the input tensor. However, in our case, the tensor describing the interactions is actually not observed. Instead, we observe the accumulation of the hidden interaction tensor, e.g. *patient-by-medication* matrix $\mathbf{M}$, where each entry $m_{ik} = c$ denotes that medication k is prescribed c times to patient i (without knowing which diagnosis it corresponds to). Note that the sum of independent Poisson distributions yields another Poisson distribution with the mean being the sum of the parameters of the composing Poisson distributions, which gives:

$$m_{ik} = \sum_{j=1}^{N_d} x_{ijk} \sim \text{Poisson}(\sum_{j=1}^{N_d} \hat{x}_{ijk}). \tag{5}$$

Together with the accumulation operation defined in Eq. (3), we can rewrite Eq. (5) in matrix form:

$$\mathbf{M} \sim \text{Poisson}(\mathbf{U}^{(1)} \,\text{diag}(\mathbf{1}^T\mathbf{U}^{(2)})\mathbf{U}^{(3)^T}). \tag{6}$$

Likewise, for the *patient-by-diagnosis* matrix $\mathbf{D}$ we have:

$$\mathbf{D} \sim \text{Poisson}(\mathbf{U}^{(1)} \,\text{diag}(\mathbf{1}^T\mathbf{U}^{(3)})\mathbf{U}^{(2)^T}). \tag{7}$$

However, for diagnoses we do not even observe the accumulation, but instead a binarized matrix $\mathbf{D}'$ with its entry d'_{ij} being one if the patient i has the diagnosis j, zero otherwise. Therefore, the elements in matrix $\mathbf{D}'$ follow a Bernoulli distribution where the probability of patient i having diagnosis k is given by:

$$\begin{aligned}\Pr(d'_{ij}=1) &= \Pr(d_{ij}>0) = 1-\Pr(d_{ij}=0)\\ &= 1-\prod_{k=1}^{N_m} e^{-\hat{x}_{ijk}}\frac{\hat{x}_{ijk}^0}{0!}\\ &= 1-\exp\left(-\sum_{r=1}^{R}\mathbf{u}_{ir}^{(1)}\left(\sum_{k=1}^{N_m}\mathbf{u}_{kr}^{(3)}\right)\mathbf{u}_{jr}^{(2)}\right).\end{aligned} \tag{8}$$

Reorganizing Eq. (8) into a more compact form, we obtain:

$$\mathbf{D}' \sim \text{Ber}\left(1-\exp\left(-\mathbf{U}^{(1)}\,\text{diag}(\mathbf{1}^T\mathbf{U}^{(3)})\mathbf{U}^{(2)^T}\right)\right). \tag{9}$$

4.2 Maximum Likelihood Estimation

The variables to be inferred are the CP factor matrices $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$ and $\mathbf{U}^{(3)}$. We derive the joint log-likelihood of observation $\mathbf{M}$ and $\mathbf{D}'$ and infer the variables by maximizing the joint log-likelihood:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}(\mathbf{M}) + \mathcal{L}(\mathbf{D}')\\ &= \sum_{i,k}\log\left(p\left(m_{ik}|\mathbf{U}^{(n)}\right)\right) + \sum_{i,j}\log\left(p\left(d'_{ij}|\mathbf{U}^{(n)}\right)\right),\end{aligned} \tag{10}$$

where the log likelihood of the *patient-by-medication* matrix $\mathbf{M}$ is given by:

$$\begin{aligned}\mathcal{L}(\mathbf{M}) &= \sum_{i,k}\log\left(p\left(m_{ik}|\mathbf{U}^{(n)}\right)\right)\\ &= \sum_{i,k}\left\{-(\sum_{j=1}^{N_d}\hat{x}_{ijk}) + m_{ik}\log(\sum_{j=1}^{N_d}\hat{x}_{ijk})\right\} + \text{constant}\\ &= \mathbf{1}^T(-\mathbf{U}^{(1)}\,\text{diag}(\mathbf{1}^T\mathbf{U}^{(2)})\mathbf{U}^{(3)^T}\\ &\quad + \mathbf{M} * \log(\mathbf{U}^{(1)}\,\text{diag}(\mathbf{1}^T\mathbf{U}^{(2)})\mathbf{U}^{(3)^T}))\mathbf{1} + \text{constant},\end{aligned} \tag{11}$$

where $*$ denotes the element-wise multiplication. The log likelihood of the *patient-by-diagnoses* matrix $\mathbf{D}'$ is given by:

$$\begin{aligned}\mathcal{L}(\mathbf{D}') &= \sum_{i,j}\log\left(p\left(d'_{ij}|\mathbf{U}^{(n)}\right)\right)\\ &= \sum_{i,j}\left(d'_{ij}\log(\exp(\sum_{k=1}^{N_m}\hat{x}_{ijk})-1) - \sum_{k=1}^{N_m}\hat{x}_{ijk}\right)\\ &= \mathbf{1}^T(\mathbf{D}' * \log(\exp(\mathbf{U}^{(1)}\,\text{diag}(\mathbf{1}^T\mathbf{U}^{(3)})\mathbf{U}^{(2)^T}) - \mathbf{E})\\ &\quad - \mathbf{U}^{(1)}\,\text{diag}(\mathbf{1}^T\mathbf{U}^{(3)})\mathbf{U}^{(2)^T})\mathbf{1}.\end{aligned} \tag{12}$$

4.3 Learning Algorithms

We estimate the variables by minimizing the negative log likelihood with non-negativity constraints. The optimization problem is formulated as follows:

$$\begin{aligned}\underset{\mathbf{U}^{(1)},\mathbf{U}^{(2)},\mathbf{U}^{(3)}}{\arg\min}\quad & f(\mathbf{U}^{(1)},\mathbf{U}^{(2)},\mathbf{U}^{(3)}) \equiv -\mathcal{L}(\mathbf{M}) - \mathcal{L}(\mathbf{D}')\\ \text{subject to}\quad & \mathbf{U}^{(n)} \geq \mathbf{0}, n = 1,2,3.\end{aligned} \tag{13}$$

We solve the optimization problem via the block coordinate descent method [Xu and Yin, 2013]. The procedure is

summarized in Algorithm 1. For each inner iteration, we fix all but one factor matrix to be updated by solving the subproblem. E.g., for the patient dimension, we have:

$$\mathbf{U}_{k+1}^{(1)} = \arg\min_{\mathbf{X}\geq\mathbf{0}} f(\mathbf{X}, \mathbf{U}_k^{(2)}, \mathbf{U}_k^{(3)}), \tag{14}$$

where the subscript k denotes the k-th iteration.

We apply the projected line search to solve the subproblem with the procedure presented in Algorithm 2. After finding the search direction $\mathbf{S}_k$, we apply the projected backtracking line search satisfying the Armijo condition [Nocedal and Wright, 2006] to ensure that the objective function decreases sufficiently in each inner iteration. Given the non-negative parameters descent step ρ ($0 < \rho < 1$) and sufficient descent σ ($0 < \sigma < 1$), we find the smallest non-negative integer t such that:

$$\begin{aligned} &f\left(P_+[\mathbf{X}_k + \rho^t \mathbf{S}_k]\right) - f\left(\mathbf{X}_k\right) \\ \leq &\sigma\left(\left(P_+[\mathbf{X}_k + \rho^t \mathbf{S}_k] - \mathbf{X}_k\right) \cdot \nabla f\left(\mathbf{X}_k\right)\right) \end{aligned} \tag{15}$$

where $P_+[\cdot]$ denotes the projection operator that projects the variable onto the feasible region, and $(\mathbf{A}\cdot\mathbf{B})$ is the inner product of matrices $\mathbf{A}$ and $\mathbf{B}$. The next iteration is then given by $\mathbf{X}_{k+1} \leftarrow P_+[\mathbf{X}_k + \rho^t \mathbf{S}_k]$. In this paper, the search direction is taken to be the negative gradient of the objective function, *i.e.*, $\mathbf{S}_k = -\nabla f(\mathbf{X}_k)$. We set $\sigma = 10^{-4}$ and $\rho = 0.5$ in the experiments.

Algorithm 1: Block Coordinate Descent Optimization Framework for HITF Model

Input : *patient-by-medication* matrix: $\mathbf{M}$, *patient-by-diagnoses* matrix: $\mathbf{D}'$
Output: CP factor matrices: $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$ and $\mathbf{U}^{(3)}$

```
initialize U^(n) (n = 1, 2, 3) randomly;
repeat
    for n = 1 : 3 do
        repeat
            update U^(n) with other variables fixed using
              projected line search in Algorithm 2;
        until subproblem converges;
    end
until all subproblems converge;
```

Algorithm 2: Projected Line Search for Solving Subproblems with Armijo Condition

Input : Variable $\mathbf{X}_k$, search direction $\mathbf{S}_k$, sufficient descent σ and descent step ρ.
Output: Updated variable $\mathbf{X}_{k+1}$

```
t ← 0 ;
while not f(P+[X_k + ρ^t S_k]) − f(X_k) ≤
  σ((P+[X_k + ρ^t S_k] − X_k) · ∇f(X_k)) do
    t ← t + 1 ;
end
update variable: X_{k+1} ← P+[X_k + ρ^t S_k];
```

Several studies have shown that zeroing out components too early is not beneficial [Ho *et al.*, 2014b; Chi and Kolda, 2012]. Thus, for practical consideration, we project the negative components to a strictly positive region $[\epsilon, +\infty)$ instead of the non-negative orthant. After convergence, we normalize the factor matrices and zeroing out the entries smaller than a threshold ϵ'. In this paper, we fix $\epsilon = \epsilon' = 10^{-5}$. Since ϵ' is chosen to be a very small positive number, omitting elements smaller than ϵ' after normalization has a very limited impact on the objective value but significantly improves the sparsity and distinction of the inferred factors.

4.4 Generalization to Higher-Order Cases

It is straightforward to generalize the proposed HITF model to the higher order cases, where N matrices $\{\mathbf{V}^{(1)}, \ldots, \mathbf{V}^{(N)}\}$ that share one dimension are given. In the higher order setting, the order of the hidden interaction tensor is $N + 1$, and the CP factor matrices are $\{\mathbf{U}^{(s)}, \mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(N)}\}$, where $\mathbf{U}^{(s)}$ is the factor matrix for the shared dimension. Then the accumulation operation for the n-th dimension defined in Eq. (3) can be generalized into the following form:

$$\hat{\mathbf{V}}^{(n)} = \mathbf{U}^{(s)} \prod_{k\neq n} \operatorname{diag}\left(\mathbf{1}^T \mathbf{U}^{(k)}\right) {\mathbf{U}^{(n)}}^T. \tag{16}$$

The log likelihood, gradient and the learning algorithm can be derived following the same strategy as described above.

The possibility of generalizing to higher-order cases enables the HITF model to consider the correspondence among patients, diagnoses, medications, lab tests and even vital signs, as the interaction of these dimensions can be of clinical importance. Another interesting observation is that the generalized formulation is similar to that proposed for Collective Matrix Factorization (CMF) [Singh and Gordon, 2008; Gunasekar *et al.*, 2016]. However, they differ in several aspects. First, CMF does not model the interaction among different dimensions, while we model the interaction explicitly. Second, in the CMF model, the reconstruction of each input matrix depends on only the shared factor matrix and the factor matrix associated with the corresponding dimension, i.e., $\hat{\mathbf{V}}^{(n)} = \mathbf{U}^{(s)}{\mathbf{U}^{(n)}}^T$. For HITF, all factor matrices contribute to the recovery of each input matrix, as shown in Eq. (16). Third, in [Gunasekar *et al.*, 2016], only the count matrices are considered but we consider the *patient-by-diagnoses* matrix $\mathbf{D}'$ being binary, which is more practical.

5 Experiments

We conduct experiments on a critical care dataset (MIMIC-III), and evaluate the quality of inferred diagnosis-medication correspondence and phenotypes. We also evaluate the accuracy of mortality prediction using the inferred phenotypes.

Data Set. MIMIC-III [Johnson *et al.*, 2016] is an open-source, large-scale, de-identified and ICU patients related EHR data set. In the MIMIC-III dataset, the patients have 11 diagnoses per visit on average. Moreover, it contains considerably many medications which are used not for treating specific diseases, such as pain relievers, making the diagnosis-medication correspondence more obscure.

Data Preprocessing. Similar to [Kim *et al.*, 2017], we extract a subset of the MIMIC-III dataset containing $7,652$ adult patients with 50% died in hospital, and only use the first admission of each patient. We exclude the base type drugs such as $D5W$ and use only the medications that appeared in at

Cardiac dysrhythmias(39.0%)		Diabetes mellitus(25.3%)		Asthma(5.5%)	
HITF	**Rubik**	**HITF**	**Rubik**	**HITF**	**Rubik**
Furosemide(0.08)	Potassium Chloride(0.08)	Insulin(0.64)	Insulin(0.09)	Albuterol 0.083% Neb Soln(0.46)	Potassium Chloride(0.08)
Potassium Chloride(0.07)	Insulin(0.06)	Insulin Human Regular(0.05)	Potassium Chloride(0.07)	Ipratropium Bromide Neb(0.39)	Insulin(0.06)
Metoprolol(0.06)	Furosemide(0.06)	Aspirin(0.05)	Furosemide(0.06)	Furosemide(0.08)	Furosemide(0.05)
Amiodarone HCl(0.05)	Magnesium Sulfate(0.04)	Furosemide(0.03)	Magnesium Sulfate(0.03)	Heparin(0.06)	Magnesium Sulfate(0.04)
Heparin Sodium(0.04)	Acetaminophen(0.03)	Atorvastatin(0.03)	Acetaminophen(0.03)		Acetaminophen(0.03)

Table 1: Top Five Corresponding Medications for Three Diagnoses Inferred by HITF and Rubik.

least 5% of the patients, resulting in 128 distinct ones. Diagnoses are grouped by the first three digits of their ICD-9 codes, and we use the diagnoses that appeared in at least 1% of the patients, which gives 184 distinct diagnoses.

Baselines. We use Rubik [Wang *et al.*, 2015], CP-APR [Chi and Kolda, 2012] and SiCNMF [Gunasekar *et al.*, 2016] as the baselines. Rubik is one of the state-of-the-art NTF-based computational phenotyping models, CP-APR is a widely used Poisson non-negative tensor factorization model and SiCNMF is based on collective matrix factorization. For Rubik and CP-APR, we adopt the two commonly used strategies to establish the interaction tensor. The first one is binary, which sets the entries of the tensor to one if the diagnosis and medication co-occur, or zero otherwise. The second one sets the tensor entries to the number of co-occurrence of medications and diagnoses.

5.1 Diagnosis-Medication Correspondence

We first evaluate the quality of the inferred correspondence. The number of phenotypes is set to 50 in this experiment.

Correspondence Inferred. To obtain the diagnosis-medication correspondence matrix of an individual patient with index i, we fix the patient index of the hidden interaction tensor at i. Since it is unrealistic to visually examine the results for all individual patients, we focus on the average correspondence over some crowds of patients. More specifically, we first select a diagnosis with index j, then extract all the patients with the selected diagnosis as the base population, and accumulate the inferred interaction tensor along patient dimension over the base population to get an average correspondence matrix. Note that in the resulting correspondence matrix, each row represents one diagnosis and each column represents a medication. We extract the j-th row and normalize it using ℓ_1 norm. The normalized value in the extracted row c_i can be interpreted as the probability that the medication i is used for treating the selected diagnosis, and we define it as the correspondence score.

Results. Due to space limitation, we select only three diagnoses and show the top five corresponding medications inferred by HITF and Rubik in Table 1. Cardiac dysrhythmias and diabetes mellitus are two common diagnoses found in the data, with 39% and 25.3% patients respectively. Asthma is a less frequent diagnosis with only 5.5% patients. The number following each medication is the correspondence score.

Diagnoses	**Medications**
Diabetes mellitus Other diseases of lung Acute kidney failure Essential hypertension ...	Insulin Insulin Human Regular
Cardiac dysrhythmias Heart failure Other diseases of lung	Amiodarone HCl Metoprolol Furosemide ...
Other diseases of lung Cardiac dysrhythmias Heart failure Chronic airway obstruction, not elsewhere classified	Albuterol Diltiazem Ipratropium Bromide MDI Fluticasone Propionate ...

Table 2: Three Examples of Inferred Phenotypes.

Table 1 suggests that Rubik fails to produce reasonable correspondence. The top corresponding medications for all the diagnoses given by Rubik are all very similar and in fact are the dominating medications in the dataset. Moreover, the correspondence score given by Rubik is not discriminative for all medications. Rubik essentially assigns most of the existing medications to every diagnosis almost evenly. On the other hand, the correspondence inferred by HITF is more reasonable. E.g., the corresponding score of insulin to diabetes inferred by HITF is 0.64, which is significantly higher than other medications. Furthermore, HITF is robust for less frequent diagnoses. Take asthma as an example. The first two corresponding medications inferred by HITF already give the overall correspondence score of 0.87, and both medications are in fact used to treat asthma in clinical practice.

Discussion. The reason for the performance improvement is that the interaction tensor in Rubik is established using the equal-correspondence strategy, and the objective function of Rubik is minimizing the reconstruction error, which makes the resulting CP factors to recover the ill-established correspondence as much as possible. In contrast, HITF only maximizes the likelihood of the observed *patient-by-medication* matrix and *patient-by-diagnosis* matrix. Under the CP factorization framework, the low-rank mechanism behind the tensor factorization leads to the discovery of more realistic and reasonable diagnosis-medication correspondence patterns.

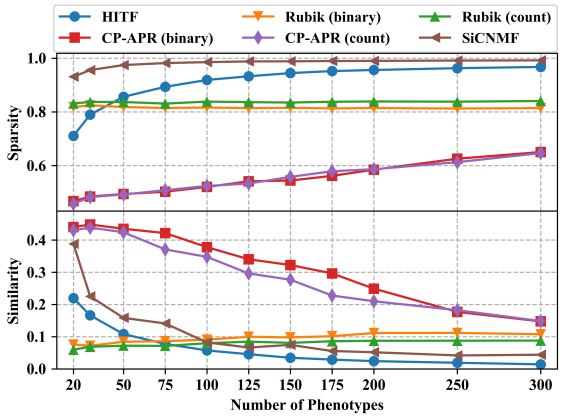

Figure 3: Performance comparison - sparsity and similarity.

5.2 Phenotypes

The clinical interpretability and meaningfulness of the derived phenotypes are important for computational phenotyping algorithms. Therefore, we first qualitatively evaluate the quality of the phenotypes derived by HITF with the number of phenotypes being set to 50. Then we increase the number of phenotypes and quantitatively measure the sparsity and similarity of the phenotypes, which are two important properties to achieve interpretability.

Qualitative Evaluation. Table 2 shows three examples of the phenotypes derived by HITF, which correspond to different patient conditions in ICU. The first one is related to patients suffering from diabetes, the second one is more related to patients having cardiac diseases, and the third one represents the patients with respiratory diseases. Note that "Other diseases of lung" is a class name in ICD-9 coding system, and most of the patients with this diagnosis is actually having acute respiratory failure, which frequently appears in ICU and can relate to many end-stage diseases.

Sparsity and Similarity. We measure the sparsity by the ratio of zero elements and the similarity by the average cosine similarity score defined as [Kim *et al.*, 2017]:

$$\text{Similarity Score} = \frac{\sum_{r_1}^{R}\sum_{r_2>r_1}^{R}\left\{\cos(\mathbf{U}^{(2)}_{:r_1},\mathbf{U}^{(2)}_{:r_2})+\cos(\mathbf{U}^{(3)}_{:r_1},\mathbf{U}^{(3)}_{:r_2})\right\}}{R(R-1)}. \quad (17)$$

where $\mathbf{U}^{(2)}, \mathbf{U}^{(3)}$ are the factor matrices associated with the diagnosis and medication dimensions respectively. We increase the number of factors from 20 to 300, and plot the sparsity and similarity scores against the number of phenotypes in Fig. 3. We see that HITF can derive significantly sparser results compared with Rubik, especially when the number of factors is large. When the number of phenotypes is set to 300, HITF derived phenotypes contain 8.5 diagnoses and 5.9 medications on average. On the other hand, when the number of phenotypes is less than 75, the phenotypes derived by HITF is less distinct than Rubik. This is because Rubik has a pairwise constraint in its objective function to ensure orthogonality in the factor matrices while HITF does not have any additive constraints on the factor matrices.

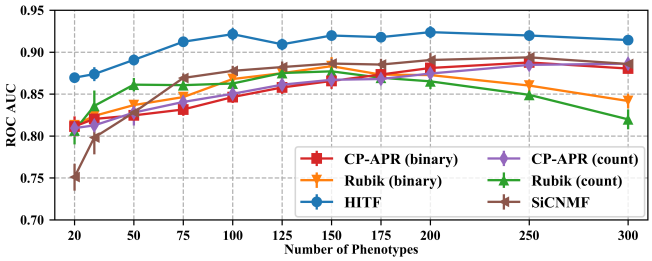

Figure 4: Prediction accuracy given different numbers of phenotypes.

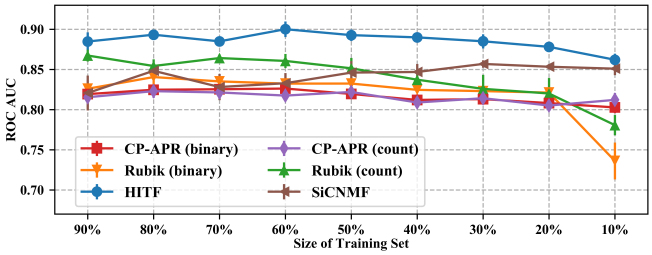

Figure 5: Prediction accuracy given different training sets.

5.3 Mortality Prediction

To examine the effectiveness of representing patients using the phenotypes derived using HITF, we apply them for in-hospital mortality prediction. We first split the data into training set and test set with a proportion of $8:2$. In the factorization step, we do not use any label information. And after the latent factors being inferred, we keep the factor matrices associated with diagnosis and medication dimensions fixed and project the test set onto the learned factors to obtain the patient representation of the test set. Then, we use a lasso regularized logistic regression to perform the binary classification. We use five-fold cross validation to train the logistic regression classier. We report the ROC AUC as a function of the number of phenotypes in Fig. 4, from which one can see HITF outperforms the baseline models significantly over all the number of phenotypes tested. In addition, as shown in Fig. 5, HITF compared with Rubik is more robust when the training set is small. The reason for HITF to outperform Rubik significantly with small training sets is that Rubik incorporates additional orthogonal constraints to enhance the interpretability, and therefore the capability of representation could be degraded given insufficient training data.

6 Conclusion

In this paper, we introduced HITF, a novel tensor factorization model to jointly learn the diagnosis-medication correspondence and phenotypes from the EHR data without the interactions among patients, medications and diagnoses being observed directly. We presented its formulation and the learning framework. One advantage of HITF compared with the existing tensor factorization models is that the hidden interactions across dimensions need not to be observed or established. Instead, HITF infers the hidden interactions together with the latent factors, making the resulting factors more accurate and precise.

The experimental results demonstrate that the diagnosis-medication correspondence learned by HITF is much more

reasonable and accurate than the equal-correspondence assumption. Moreover, the phenotypes derived by HITF are clinically meaningful and also more interpretable as they are sparser and more distinct. Furthermore, the predictive performance of HITF validates the effectiveness of representing patients using the derived phenotypes. For future research directions, we will focus on generalizing HITF to leverage multiple data sources available in ICU, such as lab test results, to discover more clinically significant patterns.

Acknowledgments

This research is partially supported by General Research Fund 12202117 from the Research Grants Council of Hong Kong.

References

[Chi and Kolda, 2012] Eric C. Chi and Tamara G. Kolda. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1272–1299, 2012.

[Gunasekar *et al.*, 2016] Suriya Gunasekar, Joyce C. Ho, Joydeep Ghosh, Stephanie Kreml, Abel N. Kho, Joshua C Denny, Bradley A Malin, and Jimeng Sun. Phenotyping using structured collective matrix factorization of multi–source EHR data. *ArXiv e-prints*, September 2016.

[Ho *et al.*, 2014a] Joyce C Ho, Joydeep Ghosh, Steve R Steinhubl, Walter F Stewart, Joshua C Denny, Bradley A Malin, and Jimeng Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics*, 52:199–211, 2014.

[Ho *et al.*, 2014b] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 115–124. ACM, 2014.

[Hripcsak and Albers, 2013] George Hripcsak and David J Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

[Jensen *et al.*, 2012] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 2016.

[Kim *et al.*, 2017] Yejin Kim, Robert El-Kareh, Jimeng Sun, Hwanjo Yu, and Xiaoqian Jiang. Discriminative and distinct phenotyping by constrained tensor factorization. *Scientific Reports*, 7(1):1114, 2017.

[Kolda and Bader, 2009] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[Lasko *et al.*, 2013] Thomas A. Lasko, Joshua C. Denny, and Mia A. Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLOS One*, 8(6):e66341, 2013.

[Liu *et al.*, 2014] Yuanyuan Liu, Fanhua Shang, Hong Cheng, James Cheng, and Hanghang Tong. Factor matrix trace norm minimization for low-rank tensor completion. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 866–874. SIAM, 2014.

[Luo *et al.*, 2016] Yuan Luo, Fei Wang, and Peter Szolovits. Tensor factorization toward precision medicine. *Briefings in Bioinformatics*, 18(3):511–514, 2016.

[Nocedal and Wright, 2006] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, 2nd edition, 2006.

[Ravì *et al.*, 2017] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21, 2017.

[Singh and Gordon, 2008] Ajit P. Singh and Geoffrey J. Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 650–658. ACM, 2008.

[Wang *et al.*, 2015] Yichen Wang, Robert Chen, Joydeep Ghosh, Joshua C Denny, Abel Kho, You Chen, Bradley A Malin, and Jimeng Sun. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1265–1274. ACM, 2015.

[Xu and Yin, 2013] Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013.

[Yamaguchi and Hayashi, 2017] Yuto Yamaguchi and Kohei Hayashi. Tensor decomposition with missing indices. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3217–3223. AAAI Press, 2017.

[Yang *et al.*, 2017] Kai Yang, Xiang Li, Haifeng Liu, Jing Mei, Guo Tong Xie, Junfeng Zhao, Bing Xie, and Fei Wang. TaGiTeD: Predictive task guided tensor decomposition for representation learning from electronic health records. In *Thirty-First AAAI Conference on Artificial Intelligence*. AAAI, 2017.

[Yu *et al.*, 2015] Sheng Yu, Katherine P Liao, Stanley Y Shaw, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Isaac S Kohane, and Tianxi Cai. Toward high-throughput phenotyping: Unbiased automated feature extraction and selection from knowledge sources. *Journal of the American Medical Informatics Association*, 22(5):993–1000, 2015.