Published in final edited form as:

Proc ACM Int Conf Inf Knowl Manag. 2019 November ; 2019: 1291–1300. doi: 10.1145/3357384.3357878.

Privacy-Preserving Tensor Factorization for Collaborative Health Data Analysis

Jing Ma,
Emory University

Qiuchen Zhang,
Emory University

Jian Lou,
Emory University

Joyce C. Ho,
Emory University

Li Xiong,
Emory University

Xiaoqian Jiang
UT Health Science Center at Houston

Abstract

Tensor factorization has been demonstrated as an efficient approach for computational phenotyping, where massive electronic health records (EHRs) are converted to concise and meaningful clinical concepts. While distributing the tensor factorization tasks to local sites can avoid direct data sharing, it still requires the exchange of intermediary results which could reveal sensitive patient information. Therefore, the challenge is how to jointly decompose the tensor under rigorous and principled privacy constraints, while still support the model's interpretability.

We propose DPFact, a privacy-preserving collaborative tensor factorization method for computational phenotyping using EHR. It embeds advanced privacy-preserving mechanisms with collaborative learning. Hospitals can keep their EHR database private but also collaboratively learn meaningful clinical concepts by sharing differentially private intermediary results. Moreover, DPFact solves the heterogeneous patient population using a structured sparsity term. In our framework, each hospital decomposes its local tensors and sends the updated intermediary results with output perturbation every several iterations to a semi-trusted server which generates the phenotypes. The evaluation on both real-world and synthetic datasets demonstrated that under strict privacy constraints, our method is more accurate and communication-efficient than state-of-the-art baseline methods.

jing.ma@emory.edu.

Keywords

Phenotyping; Tensor Factorization; Collaborative Learning; Differential Privacy

1 INTRODUCTION

Electronic Health Records (EHRs) have become an important source of comprehensive information for patients' clinical histories. While EHR data can help advance biomedical discovery, this requires an efficient conversion of the data to succinct and meaningful patient characterizations. Computational phenotyping is the process of transforming the noisy, massive EHR data into meaningful medical concepts that can be used to predict the risk of disease for an individual, or the response to drug therapy. Phenotyping can be used to assist precision medicine, speedup biomedical discovery, and improve healthcare quality [25, 29].

Yet, extracting precise and meaningful phenotypes from EHRs is challenging because observations in EHRs are high-dimensional and heterogeneous, which leads to poor interpretability and research quality for scientists [29]. Traditional phenotyping approaches require the involvement of medical domain experts, which is time-consuming and labor-intensive. Recently, unsupervised learning methods have been demonstrated as a more efficient approach for computational phenotyping. Although these methods do not require experts to manually label the data, they require large volumes of EHR data. A popular unsupervised phenotyping approach is tensor factorization [15, 20, 28]. Not only can tensors capture the interactions between multiple sources (e.g, specific procedures that are used to treat a disease), it can identify patient subgroups and extract concise and potentially more interpretable results by utilizing the multi-way structure of a tensor.

However, one existing barrier for high-throughput tensor factorization is that EHRs are fragmented and distributed among independent medical institutions, where healthcare practises are different due to heterogeneous patients populations. One of the reasons is that different hospitals or medical sites differ in the way they manage patients [31]. Moreover, effective phenotyping requires a large amount of data to guarantee its reliance and generalizability. Simply analyzing data from single source leads to poor accuracy and bias, which would reduce the quality and efficiency of patients' care.

Recent studies have suggested that the integration of health records can provide more benefits [12], which motivated the application of federated tensor learning framework [20]. It can mitigate privacy issues under the distributed data setting while achieves high global accuracy and data harmonization via federated computation. But this method has inherent limitations of federated learning: 1) high communication cost; 2) reduced accuracy due to local non-IID data (i.e., patient heterogeneity); and 3) no formal privacy guarantee of the intermediary results shared between local sites and the server, which makes patient data at risk of leakage.

In this paper, we propose DPFact, a differentially private collaborative tensor factorization framework based on Elastic Averaging Stochastic Gradient Descent (EASGD) for computational phenotyping. DPFact assumes all sites share a common model learnt jointly

from each site through communication with a central parameter server. Each site performs its own tensor factorization task to discover both common and distinct latent components, while benefiting from the intermediary results generated by other sites. The intermediary results uploaded still contain sensitive information about the patients. Several studies have shown that machine learning models can be used to extract sensitive information used in the input training data through membership inference attacks or model inversion attacks both in the centralized setting [11, 26] and federated setting [14]. Since we assume the central server and participants are honest-but-curious, hence a formal differential privacy guarantee is desired. DPFact tackles the privacy issue with a well-designed data-sharing strategy, combined with the rigorous zero-concentrated differential privacy (zCDP) technique [9, 34] which is a strictly stronger definition than (ϵ, δ)-differential privacy that is considered as the dominant standard for strong privacy protection [8–10]. We briefly summarize our contributions as:

1) **Efficiency.** DPFact achieves higher accuracy and faster convergence rate than the state-of-the-art federated learning method. It also beats the federated learning method in achieving lower communication cost thanks to the elimination of auxiliary parameters (e.g., in the ADMM approach) and allows each local site to perform most of the computation.

2) **Utility.** DPFact supports phenotype discovery even with a rigorous privacy guarantee. By incorporating a $l_{2,1}$ regularization term, DPFact can jointly decompose local tensors with different distribution patterns and discover both the globally shared and the distinct, site-specific phenotypes.

3) **Privacy.** DPFact is a privacy-preserving collaborative tensor factorization framework. By applying zCDP mechanisms, it guarantees that there is no inadvertent patient information leakage in the process of intermediary results exchange with high probability which is quantified by privacy parameters.

We evaluate DPFact on two publicly-available large EHR datasets and a synthetic dataset. The performance of DPFact is assessed from the following three aspects including efficiency measured by accuracy and communication cost, utility measured by phenotype discovery ability and the evaluation on the effect of privacy.

2 PRELIMINARIES AND NOTATIONS

This section describes the preliminaries used in this paper, including tensor factorization, (ϵ, δ)-differential privacy, and zCDP.

2.1 Tensor Factorization

***Definition 2.1.* (*Khatri-Rao product*).—**Khatri-Rao product is the "columnwise" Kronecker product of two matrices $\mathbf{A} \in \mathbb{R}^{I \times R}$ and $\mathbf{B} \in \mathbb{R}^{J \times R}$. The result is a matrix of size ($IJ \times R$) and defined by

$$\mathbf{A} \odot \mathbf{B} = \left[\mathbf{a}_1 \otimes \mathbf{b}_1 \cdots \mathbf{a}_R \otimes \mathbf{b}_R\right]$$

Here, $\otimes$ denotes the *Kronecker product*. The *Kronecker product* of two vectors, $\mathbf{a} \in \mathbb{R}^I$ and $\mathbf{b} \in \mathbb{R}^J$ is

$$\mathbf{a} \otimes \mathbf{b} = \begin{bmatrix} a_1\mathbf{b} \\ \vdots \\ a_I\mathbf{b} \end{bmatrix}$$

***Definition 2.2.* (*CANDECOMP-PARAFAC Decomposition*).—**The CANDECOMP-PARAFAC (CP) decomposition is to approximate the original tensor $\mathcal{O}$ by the sum of R rank-one tensors. R is the rank of tensor $\mathcal{O}$, It can be expressed as

$$\mathcal{O} \approx \mathcal{X} = \sum_{r=1}^{R} \mathbf{a}_{:r}^{(1)} \circ \cdots \circ \mathbf{a}_{:r}^{(N)}, \tag{1}$$

where $\mathbf{a}_{:r}^{(n)}$ represents the r^{th} column of $A^{(n)}$ for $n = 1, \cdots, N$ and $r = 1, \cdots, R$. $A^{(n)}$ is the n-mode factor matrix consisting of R columns representing R latent components which can be represented as

$$A^{(n)} = \left[\mathbf{a}_{:1}^{(n)} \cdots \mathbf{a}_{:R}^{(n)}\right],$$

so that $A^{(n)}$ is of size $I_n \times R$ for $n = 1, \cdots, N$, and the equation of (1) can also be represented as

$$[\![A^{(1)}, \cdots, A^{(N)}]\!] = \sum_{r=1}^{R} \mathbf{a}_{:r}^{(1)} \circ \cdots \circ \mathbf{a}_{:r}^{(N)}. \tag{2}$$

Note that in this formulation, the scalar weights for each rank-one tensor are assumed to be absorbed into the factors.

In the way of a three-mode tensor $\mathcal{O} \in \mathbb{R}^{I \times J \times K}$, the CP decomposition can be represented as

$$\mathcal{O} \approx \mathcal{X} = \sum_{r=1}^{R} \mathbf{a}_{:r} \circ \mathbf{b}_{:r} \circ \mathbf{c}_{:r}, \tag{3}$$

where $\mathbf{a}_{:r} \in \mathbb{R}^I$, $\mathbf{b}_{:r} \in \mathbb{R}^J$, $\mathbf{c}_{:r} \in \mathbb{R}^K$ are the r-th column vectors within the three factor matrices $\mathbf{A} \in \mathbb{R}^{I \times R}, \mathbf{B} \in \mathbb{R}^{J \times R}, \mathbf{C} \in \mathbb{R}^{K \times R}$.

2.2 Differential Privacy

Differential privacy [8, 9] has been demonstrated as a strong standard to provide privacy guarantees for algorithms on aggregate database analysis, which in our case is a

collaborative tensor factorization algorithm analyzing distributed tensors with differential privacy.

Definition 2.3. **((ϵ-δ)-Differential Privacy) [8].—**Let $\mathscr{D}$ and $\mathscr{D}'$ be two neighboring datasets that differ in at most one entry. A randomized algorithm $\mathscr{A}$ is (ϵ-δ)-differentially private if for all $\mathscr{S} \subseteq \text{Range}(\mathscr{A})$:

$$\Pr[\mathscr{A}(\mathscr{D}) \in \mathscr{S}] \leq e^{\epsilon} Pr[\mathscr{A}(\mathscr{D}') \in \mathscr{S}] + \delta,$$

where $\mathscr{A}(\mathscr{D})$ represents the output of $\mathscr{A}$ with an input of $\mathscr{D}$.

The above definition suggests that with a small ϵ, an adversary almost cannot distinguish the outputs of an algorithm with two neighboring datasets $\mathscr{D}$ and $\mathscr{D}'$ as its inputs. While δ allows a small probability of failing to provide this guarantee. Differential privacy is defined using a pair of neighboring databases which in our work are two tensors and differ in only one entry.

Definition 2.4. **(L_2-sensitivity) [8].—**For two neighboring datasets $\mathscr{D}$ and $\mathscr{D}'$ differing in at most one entry, the L_2-sensitivity of an algorithm $\mathscr{D}$ is the maximum change in the l_2-norm of the output value of algorithm $\mathscr{D}$ regarding the two neighboring datasets:

$$\Delta_2(\mathscr{A}) = \sup_{\mathscr{D}, \mathscr{D}'} \|\mathscr{A}(\mathscr{D}) - \mathscr{A}(\mathscr{D}')\|_2.$$

THEOREM 2.5. ***((Gaussian Mechanism)) [8].—****Let* $\epsilon \in (0, 1)$ *be arbitrary. For* $c^2 > 2 \ln(1.25/\delta)$, *the Gaussian Mechanism with parameter* $\sigma \geq c\Delta_2(\mathscr{A})/\epsilon$, *adding noise scaled to* $\mathcal{N}(0, \sigma^2)$ *to each component of the output of algorithm* $\mathscr{A}$, *is* (ϵ-δ)*-differentially private.*

2.3 Concentrated Differential Privacy

Concentrated differential privacy (CDP) is introduced by Dwork and Rothblum [9] as a generalization of differential privacy which provides sharper analysis of many privacy-preserving computations. Bun and Steinke [4] propose an alternative formulation of CDP called “zero-concentrated differential privacy” (zCDP) which utilizes the Rényi divergence between probability distributions to measure the requirement of the privacy loss random variable to be sub-gaussian and provides more tighter privacy analysis.

Definition 2.6. **(*Zero-Concentrated Differential Privacy (zCDP) [4]*)—**A randomized mechanism $\mathscr{A}$ is ρ-zero concentrated differentially private if for any two neighboring databases $\mathscr{D}$ and $\mathscr{D}'$ differing in at most one entry and all $\alpha \in (1, \infty)$,

$$D_\alpha(\mathscr{A}(\mathscr{D}) \| \mathscr{A}(\mathscr{D}')) \triangleq \frac{1}{\alpha - 1} \log\left(\mathbb{E}\left[e^{(\alpha - 1)L^{(o)}}\right]\right) \leq \rho\alpha,$$

where $D_\alpha(\mathscr{A}(\mathscr{D})\|\mathscr{A}(\mathscr{D}'))$ is called α-Rényi divergence between the distributions of $\mathscr{A}(\mathscr{D})$ and $\mathscr{A}(\mathscr{D}')$, and $L(o)$ is the privacy loss random variable which is defined as:

$$L^{o}_{(\mathscr{A}(\mathscr{D})\|\mathscr{A}(\mathscr{D}'))} \triangleq \log\frac{\Pr(\mathscr{A}(\mathscr{D})=o)}{\Pr(\mathscr{A}(\mathscr{D}')=o)}.$$

The following propositions of zCDP will be used in this paper.

PROPOSITION 2.7.—*[4] The Gaussian mechanism with noise* $\mathscr{N}(0, \sigma^2)$ *where* $\sigma=\sqrt{1/(2\rho)}\Delta_2$ *satisfies* ρ*-zCDP.*

PROPOSITION 2.8.—*[4] If a randomized mechanism* $\mathscr{A}$ *is* ρ*-CDP, then* $\mathscr{A}$ *is* (ϵ', δ)*-DP for any* δ *with* $\epsilon'=\rho+\sqrt{4\rho\log(1/\delta)}$*; For* $\mathscr{A}$ *to satisfy* (ϵ, δ)*-DP, it suffices to satisfy* ρ*-zCDP by setting* $\rho\approx\frac{\epsilon^2}{4\log(1/\delta)}$.

PROPOSITION 2.9.—*((Serial composition [4])) Let* $\mathscr{A}:\mathscr{D}^n\to\mathscr{Y}$ *and* $\mathscr{A}':\mathscr{D}^n\to\mathscr{Z}$ *be randomized algorithms. Suppose* $\mathscr{A}$ *is* ρ*-zCDP and* $\mathscr{A}'$ *is* ρ'*-zCDP. Define* $\mathscr{A}'':\mathscr{D}^n\to\mathscr{Y}\times\mathscr{Z}$ *by* $\mathscr{A}''=(\mathscr{A},\mathscr{A}')$*. Then* $\mathscr{A}''$ *is* $(\rho+\rho')-zCDP$

PROPOSITION 2.10.—*((Parallel composition [34])) Suppose that a mechanism* $\mathscr{A}$ *consists of a sequence of T adaptive mechanisms,* $\mathscr{A}_1,\ldots,\mathscr{A}_T$*, where each* $\mathscr{A}_t:\prod_{j=1}^{iter-1}\mathscr{O}_j\times\mathscr{D}_t\to\mathscr{O}_{iter}$ *and* $\mathscr{A}_t$ *satisfies* ρ_t*-zCDP. Let* $\mathscr{D}_1,\ldots,\mathscr{D}_T$ *be a randomized partition of the input* $\mathscr{D}$*. The mechanism* $\mathscr{A}(\mathscr{D})=\left(\mathscr{A}_1(\mathscr{D}_1),\ldots,\mathscr{A}_T(\mathscr{D}_T)\right)$ *satisfies* $\frac{1}{T}\sum_{t=1}^{T}\rho_t-zCDP$.

3 DPFACT

In this section, we first provide a general overview and then present detailed formulation of the optimization problem.

3.1 Overview

DPFact is a distributed tensor factorization model that preserves differential privacy. Our goal is to learn computational phenotypes from horizontally partitioned patient data (e.g., each hospital has its own patient data with the same medical features). Since we assume the central server and participants are honest-but-curious which means they will not deviate from the prescribed protocol but they are curious about others secrets and try to find out as much as possible about them. Therefore the patient data cannot be collected at a centralized location to construct a global tensor $\mathscr{O}$. Instead, we assume that there are T local sites and a central server that communicates the intermediary results between the local sites. Each site performs tensor factorization on the local data and shares privacy-preserving intermediary results with the centralized server (Figure 1).

The patient data at each site is used to construct a local observed tensor, $\mathcal{O}^{[t]}$. For simplicity and illustration purposes, we discuss a three-mode tensor situation where the modes are patients, procedures, and diagnoses but DPFact generalizes to N modes. The T sites jointly decompose their local tensor into three factor matrices: a patient factor matrix $\mathbf{A}^{[t]}$ and two feature factor matrices $\mathbf{B}^{[t]}$ and $\mathbf{C}^{[t]}$. We assume that the factor matrices on the non-patient modes (i.e., $\mathbf{B}^{[t]}$, $\mathbf{C}^{[t]}$) are the same across the T sites, thus sharing the same computational phenotypes. To achieve consensus of the shared factor matrices, the non-patient feature factor matrices are shared in a privacy-preserving manner with the central server by adding Gaussian noise to each uploaded factor matrix.

Although the collaborative tensor problem for computational phenotyping has been previously discussed [20], DPFact provides three important contributions:

(1) **Efficiency:** We adopt a communication-efficient stochastic gradient descent (SGD) algorithm for collaborative learning which allows each site to transmit less information to the centralized server while still achieving an accurate decomposition.

(2) **Heterogeneity:** A traditional global consensus model requires learning the same shared model from multiple sources. However, different data sources may have distinct patterns and properties (e.g., disease prevalence may differ between Georgia and Texas). We propose using the $l_{2,1}$-norm to achieve global consensus among the sites while capturing site-specific factors.

(3) **Differential Privacy Guarantees:** We preserve the privacy of intermediary results by adding Gaussian noise to each non-patient factor matrix prior to sharing with the parameter server. This masks any particular entry in the factor matrices and prevents inadvertent privacy leakage. A rigorous privacy analysis based on zCDP is performed to ensure strong privacy protection for the patients.

3.2 Formulation

Under a single (centralized) model, CP decomposition of the observed tensor $\mathcal{O}$ results in a factorized tensor $\mathcal{X}$ that contains the R most prevalent computational phenotypes. We represent the centralized tensor as T separate horizontal partitions, $\mathcal{O}^{[1]}, \cdots, \mathcal{O}^{[T]}$. Thus, the global function can be expressed as the sum of T separable functions with respect to each local factorized tensor $\mathcal{X}^{[t]}$ [20]:

$$\min_{\mathcal{X}} \mathcal{L} = \frac{1}{2}\|\mathcal{O} - \mathcal{X}\|_F^2 = \sum_{t=1}^{T} \frac{1}{2}\left\|\mathcal{O}^{[t]} - \mathcal{X}^{[t]}\right\|_F^2. \quad (4)$$

Since the goal is to uncover computational phenotypes that are shared across all sites, we restrict the sites to factorize the observed local tensors $\mathcal{O}^{[t]}$ such that the non-patient factor matrices are the same. Therefore, the global optimization problem is formulated as:

$$\min \sum_{t=1}^{T} \frac{1}{2} \| \mathcal{O}^{[t]} - [\![\mathbf{A}^{[t]}, \mathbf{B}^{[t]}, \mathbf{C}^{[t]}]\!] \|_F^2$$

$$\text{s.t. } \mathbf{B}^{[1]} = \mathbf{B}^{[2]} = \cdots = \mathbf{B}^{[T]}$$

$$\mathbf{C}^{[1]} = \mathbf{C}^{[2]} = \cdots = \mathbf{C}^{[T]}.$$

This can be reformulated as a global consensus optimization, which decomposes the original problem into T local subproblems by introducing two auxiliary variables, $\widehat{\mathbf{B}}$, $\widehat{\mathbf{C}}$, to represent the global factor matrices. A quadratic penalty is placed between the local and global factor matrices to achieve global consensus among the T different sites. Thus, the local optimization problem at site t is:

$$\begin{aligned} \min & \frac{1}{2} \| \mathcal{O}^{[t]} - [\![\mathbf{A}^{[t]}, \mathbf{B}^{[t]}, \mathbf{C}^{[t]}]\!] \|_F^2 \\ & + \frac{\gamma}{2} \| \mathbf{B}^{[t]} - \widehat{\mathbf{B}} \|_F^2 + \frac{\gamma}{2} \| \mathbf{C}^{[t]} - \widehat{\mathbf{C}} \|_F^2. \end{aligned} \tag{5}$$

3.3 Heterogeneous Patient Populations

The global consensus model assumes that the patient populations are the same across different sites. However, this may be too restrictive as some locations can have distinctive patterns. For example, patients from the cardiac coronary unit may have unique characteristics that are different from the surgical care unit. DPFact utilizes the $l_{2,1}$-norm regularization, to allow flexibility for each site to "turn off" one or more computational phenotypes. For an arbitrary matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$, its $l_{2,1}$-norm is defined as:

$$\| \mathbf{W} \|_{2,1} = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} \mathbf{W}_{ij}^2}. \tag{6}$$

From the definition, we can see that the $l_{2,1}$-norm controls the row sparsity of matrix $\mathbf{W}$. As a result, the $l_{2,1}$-norm is commonly used in multi-task feature learning to perform feature selection as it can induce structural sparsity [13, 21, 24, 32].

DPFact adopts a multi-task perspective, where each local decomposition is viewed as a separate task. Under this approach, each site is not required to be characterized by all R computational phenotypes. To achieve this, we introduce the $l_{2,1}$-norm on the transpose of the patient factor matrices, $\mathbf{A}^{[t]}$, to induce sparsity on the columns. The idea is that if a specific phenotype is barely present in any of the patients (2-norm of the column is close to 0), the regularization will encourage all the column entries to be 0. This can be used to capture the heterogeneity in the patient populations without violating the global consensus assumption. Thus the DPFact optimization problem is:

$$\min \sum_{t=1}^{T} \left(\frac{1}{2} \| \mathcal{O}^{[t]} - [\![\mathbf{A}^{[t]}, \mathbf{B}^{[t]}, \mathbf{C}^{[t]}]\!] \|_F^2 + \frac{\gamma}{2} \| \mathbf{B}^{[t]} - \widehat{\mathbf{B}} \|_F^2 + \frac{\gamma}{2} \| \mathbf{C}^{[t]} - \widehat{\mathbf{C}} \|_F^2 + \mu \| (\mathbf{A}^{[t]})^\top \|_{2,1} \right). \tag{7}$$

The quadratic penalty, γ, provides an elastic force to achieve global consensus between the local factor matrices and the global factor matrices whereas the $l_{2,1}$-norm penalty, μ, encourages sites to share similar sparsity patterns.

4 DPFACT OPTIMIZATION

DPFact adopts the Elastics Averaging SGD (EASGD) [35] approach to solve the optimization problem (7). EASGD is a communication-efficient algorithm for collaborative learning and has been shown to be more stable than the Alternating Direction Method of Multipliers (ADMM) with regard to parameter selection. Moreover, SGD-based approaches scale well to sparse tensors, as the computation is bounded by the number of non-zeros.

Using the EASGD approach, the global consensus optimization problem is solved alternatively between the local sites and the central server. Each site performs multiple rounds of local tensor decomposition and updates their local factor matrices. The site then only shares the most updated non-patient mode matrices with output perturbation to prevent revealing of sensitive information. The patient factor matrix is never shared with the central server to avoid direct leakage of patient membership information. The server then aggregates the updated local factor matrices to update the global factor matrices and sends the new global factor matrices back to each site. This process is iteratively repeated until there are no changes in the local factor matrices. The entire DPFact decomposition process is summarized in Algorithm 1.

Input: O, τ η, γ, μ, σ, ρ.
Randomly initialize the global feature factor matrices $\mathbf{B}$, $\mathbf{C}$ and local feature factor matrices $\mathbf{B}^{[t]}$, $\mathbf{C}^{[t]}$.
while $B^{[t]}$, $C^{[t]}$ *not converge* **do**
  if ***Hospital*** **then**
    for $k = 1, \cdots, \tau$ **do**
      Shuffle tensor elements;
      for *observation i* **do**
        Update $\mathbf{A}^{[t]}$ using (13);
        Update $\mathbf{B}^{[t]}$, $\mathbf{C}^{[t]}$ using (17);
      end
      Proximal update for ${}_{new}\mathbf{A}^{[t]}$ using (14);
    end
    Calibrate *Gaussian* noise matrix $\mathcal{M}_B^{[t]}$ and $\mathcal{M}_C^{[t]}$ as $\mathcal{N}(0, \Delta_2^2/(2\rho))$ for each factor matrix;
    Update factor matrices ${}_{priv}\mathbf{B}^{[t]}$ and ${}_{priv}\mathbf{C}^{[t]}$ using (18);
    Send ${}_{priv}\mathbf{B}^{[t]}$, ${}_{priv}\mathbf{C}^{[t]}$ to Server.
  end
  if ***Server*** **then**
    Collect ${}_{priv}\mathbf{B}^{[t]}$, ${}_{priv}\mathbf{C}^{[t]}$ from each hospital;
    Update $\widehat{\mathbf{B}}$, $\widehat{\mathbf{C}}$ using (19);
    Send $\widehat{\mathbf{B}}$, $\widehat{\mathbf{C}}$ back to hospitals.
  end
end

4.1 Local Factors Update

Each site updates the local factors by solving the following subproblem:

$$\min \frac{1}{2}\|\mathcal{O}^{[t]} - [\![\mathbf{A}^{[t]}, \mathbf{B}^{[t]}, \mathbf{C}^{[t]}]\!] \|_F^2 + \frac{\gamma}{2}\|\mathbf{B}^{[t]} - \widehat{\mathbf{B}}\|_F^2 + \frac{\gamma}{2}\|\mathbf{C}^{[t]} - \widehat{\mathbf{C}}\|_F^2 + \mu\|(\mathbf{A}^{[t]})^\top\|_{2,1}. \tag{8}$$

EASGD helps reduce the communication cost by allowing sites to perform multiple iterations (each iteration is one pass of the local data) before sending the updated factor matrices. We further extend the local optimization updates using permutation-based SGD (P-SGD), a practical form of SGD [30]. In P-SGD, instead of randomly sampling one instance from the tensor at a time, the non-zero elements are first shuffled within the tensor. The algorithm then cycles through these elements to update the latent factors. At each local site, the shuffling and cycling process is repeated τ times, hereby referred to as a τ-pass P-SGD. There are two benefits of adopting the P-SGD approach: 1) the resulting algorithm is more computationally effective as it eliminates some of the randomness of the basic SGD algorithm. 2) it provides a mechanism to properly estimate the total privacy budget (see Section 4.2).

4.1.1 Patient Factor Matrix.—

For site t, the patient factor matrix $\mathbf{A}^{[t]}$ is updated by minimizing the objective function using the local factorized tensor, $\mathcal{X}^{[t]}$ and the $l_{2,1}$-norm:

$$\min_{\mathbf{A}^{[t]}} \underbrace{\frac{1}{2}\|\mathcal{O}^{[t]} - [\![\mathbf{A}^{[t]}, \mathbf{B}^{[t]}, \mathbf{C}^{[t]}]\!] \|_F^2}_{\mathscr{F}} + \underbrace{\mu\|(\mathbf{A}^{[t]})^\top\|_{2,1}}_{\mathscr{H}}. \tag{9}$$

While the $l_{2,1}$-norm is desirable from a modeling perspective, it also results in a non-differentiable optimization problem. The local optimization problem (9) can be seen as a combination of a differentiable function $\mathscr{F}$ and a non-differentiable function $\mathscr{H}$. Thus, we propose using the proximal gradient descent method to solve local optimization problem for the patient mode. Proximal gradient method can be applied in our case since the gradient of the differentiable function $\mathscr{F}$ is *Lipschitz* continuous with a *Lipschitz* constant L (see Appendix in [23] for details).

Using the proximal gradient method, the factor matrix $\mathbf{A}^{[t]}$ is iteratively updated via the proximal operator:

$${}_{new}\mathbf{A}^{[t]} = \mathbf{prox}_{\eta\mathscr{H}}\left(\mathbf{A}^{[t]} - \eta\nabla\mathscr{F}(\mathbf{A}^{[t]})\right), \tag{10}$$

where $\eta > 0$ is the step size at each local iteration. The proximal operator is computed by solving the following equation:

$$\mathbf{prox}_{\eta\mathscr{H}}(\Theta) = \underset{\Theta}{\operatorname{argmin}}\left(\frac{1}{2\eta}\|\Theta - \widehat{\Theta}\| + \mathscr{H}(\Theta)\right), \tag{11}$$

where $\widehat{\Theta} = \mathbf{A}^{[t]} - \eta\nabla\mathscr{F}(\mathbf{A}^{[t]})$ is the updated matrix. It has been shown that if $\nabla\mathscr{F}$ is *Lipschitz* continuous with constant L, the proximal gradient descent method will converge for step size

$\eta < 2/L$ [7]. For the $l_{2,1}$-norm, the closed form solution can be computed using the soft-thresholding operator:

$$\mathbf{prox}_{\eta\mathscr{R}}(\widehat{\Theta}) = \widehat{\Theta}_{:r}\left(1 - \frac{\mu}{\|\widehat{\Theta}_{:r}\|_2}\right)_+, \tag{12}$$

where $r \in (0, R]$ and r represents the r-th column of the factor matrix $\widehat{\Theta}$, and $(z)_+$ denotes the maximum of 0 and z. Thus, if the norm of the r-th column of the patient matrix is small, the proximal operator will "turn off" that column.

The gradient of the smooth part can be derived with respect to each row in the patient mode factor matrix, $\mathbf{A}^{[t]}$. The update rule for each row is:

$$\mathbf{a}_{i:}^{[t]} \leftarrow \mathbf{a}_{i:}^{[t]} - \eta\left[\left(\mathbf{a}_{i:}^{[t]}(\mathbf{b}_{j:}^{[t]} * \mathbf{c}_{k:}^{[t]})^\top - \mathcal{O}_{ijk}^{[t]}\right)\left(\mathbf{b}_{j:}^{[t]} * \mathbf{c}_{k:}^{[t]}\right)\right] \tag{13}$$

After one pass through all entries in a local tensor to update the patient factor matrix, the second step is to use proximal operator (12) to update the patient factor matrix $\mathbf{A}^{[t]}$:

$$_{new}\mathbf{A}^{[t]} = \mathbf{prox}_{\eta\mathscr{R}}(\mathbf{A}^{[t]}). \tag{14}$$

4.1.2 Feature Factor Matrices.—

The local feature factor matrices, $\mathbf{B}^{[t]}$ and $\mathbf{C}^{[t]}$, are updated based on the following objective functions:

$$\begin{aligned}
\min_{\mathbf{B}^{[t]}} f_b &= \frac{1}{2}\|\mathcal{O}^{[t]} - [\![\, \mathbf{A}^{[t]}, \mathbf{B}^{[t]}, \mathbf{C}^{[t]} \,]\!]\,\|_F^2 + \frac{\gamma}{2}\|\mathbf{B}^{[t]} - \widehat{\mathbf{B}}\|_F^2, \\
\min_{\mathbf{C}^{[t]}} f_c &= \frac{1}{2}\|\mathcal{O}^{[t]} - [\![\, \mathbf{A}^{[t]}, \mathbf{B}^{[t]}, \mathbf{C}^{[t]} \,]\!]\,\|_F^2 + \frac{\gamma}{2}\|\mathbf{C}^{[t]} - \widehat{\mathbf{C}}\|_F^2.
\end{aligned} \tag{15}$$

The partial derivatives of f_b, f_c with respect to $\mathbf{b}_{j:}^{[t]}$ and $\mathbf{c}_{k:}^{[t]}$, the j-th and k-th row of the $\mathbf{B}^{[t]}$ and $\mathbf{C}^{[t]}$ factor matrices, respectively, are computed.

$$\begin{aligned}
\frac{\partial f_b}{\partial \mathbf{b}_{j:}^{[t]}} &= \left[\left(\mathbf{a}_{i:}^{[t]}(\mathbf{b}_{j:}^{[t]} * \mathbf{c}_{k:}^{[t]})^\top - \mathcal{O}_{ijk}^{[t]}\right)\left(\mathbf{a}_{i:}^{[t]} * \mathbf{c}_{k:}^{[t]}\right)\right] \\
\frac{\partial f_c}{\partial \mathbf{c}_{k}^{[t]}} &= \left[\left(\mathbf{a}_{i:}^{[t]}(\mathbf{b}_{j:}^{[t]} * \mathbf{c}_{k:}^{[t]})^\top - \mathcal{O}_{ijk}^{[t]}\right)\left(\mathbf{a}_{i:}^{[t]} * \mathbf{b}_{j:}^{[t]}\right)\right].
\end{aligned} \tag{16}$$

$\mathbf{B}^{[t]}$ and $\mathbf{C}^{[t]}$ are then updated row by row by adding up the partial derivative of the quadratic penalty term and the partial derivative with respect to $\mathbf{b}_{j:}^{[t]}$ and $\mathbf{c}_{k:}^{[t]}$ shown in (16).

$$\mathbf{b}_{j:}^{[t]} \leftarrow \mathbf{b}_{j:}^{[t]} - \eta\left[\frac{\partial f_n}{\partial \mathbf{b}_j^{[t]}} + \gamma\left(\mathbf{b}_{j:}^{[t]} - \widehat{\mathbf{b}_{j:}}\right)\right];$$
$$\mathbf{c}_{k:}^{[t]} \leftarrow \mathbf{c}_{k:}^{[t]} - \eta\left[\frac{\partial f_n}{\partial \mathbf{c}_k^{[t]}} + \gamma\left(\mathbf{c}_{k:}^{[t]} - \widehat{\mathbf{c}_{k:}}\right)\right]. \quad (17)$$

Each site simultaneously does several rounds (τ) of the local factor updates. After τ rounds are completed, the feature factor matrices will be perturbed with *Gaussian* noise and sent to central server.

4.1.3 Privacy-Preserving Output Perturbation.—

Although the feature factor matrices do not directly contain patient information, it may inadvertently violate patient privacy (e.g., a rare disease that is only present in a small number of patients). To protect the patient information from being speculated by semi-honest server, we perturb the feature mode factor matrices using the Gaussian mechanism, a common building block to perturb the output and achieve rigorous differential privacy guarantee.

The Gaussian mechanism adds zero-mean Gaussian noise with standard deviation $\sigma = \Delta_2^2/(2\rho)$ to each element of the output [4]. Thus, the noise matrix $\mathscr{M}$ can be calibrated for each factor matrices $\mathbf{B}^{[t]}$ and $\mathbf{C}^{[t]}$ based on their L_2-sensitivity to construct privacy-preserving feature factor matrices:

$$_{priv}\mathbf{B}^{[t]} \leftarrow \mathbf{B}^{[t]} + \mathscr{M}_B^{[t]},$$
$$_{priv}\mathbf{C}^{[t]} \leftarrow \mathbf{C}^{[t]} + \mathscr{M}_C^{[t]}, \quad (18)$$

As a result, each factor matrix that is shared with the central server satisfies ρ-zCDP by Proposition 2.7. A detailed privacy analysis for the overall privacy guarentee is provided in the next subsection.

4.2 Privacy Analysis

In this subsection we analyze the overall privacy guarantee of Algorithm 1. The analysis is based on the following knowledge of the optimization problem: 1) each local site performs a τ-pass P-SGD update per epoch; 2) for the local objective function f in (15), when fixing two of the factor matrices, the objective function becomes a convex optimization problem for the other factor matrix.

4.2.1 L_2-sensitivity.—

The objective function (15) satisfies L – *Lipschitz*, with *Lipschitz* constant L the tight upper bound of the gradient of $\mathbf{B}^{[t]}$ and $\mathbf{C}^{[t]}$. For a τ-pass P-SGD, having constant learning rate $\eta = \eta_k \leq \frac{2}{\beta}(k = 1, \ldots, \tau, \beta$ is the *Lipschitz* constant of the gradient of (15) regarding $\mathbf{B}^{[t]}$ or $\mathbf{C}^{[t]}$, see Appendix in [23] for β calculation), the L_2-sensitivity of this optimization problem in (15) is calculated as $\Delta_2(f) = 2\tau L\eta$ [30].

4.2.2 Overall Privacy Guarantee.—The overall privacy guarantee of Algorithm 1 is analyzed under the zCDP definition which provides tighter privacy bound than strong composition theorem [10] for multiple folds Gaussian mechanism [4, 34]. The total ρ-zCDP will be transferred to (ϵ, δ)-DP in the end using Proposition 2.8.

THEOREM 4.1.: *Algorithm 1 is* (ϵ, δ)*-differentially private if we choose the input privacy budget for each factor matrix per epoch as*

$$\rho = \frac{\epsilon^2}{8E\log(1/\delta)}$$

where E is the number of epochs when the algorithm is converged.

PROOF.: Let the "base" zCDP parameter be ρ_b, $\mathbf{B}^{[t]}$ and $\mathbf{C}^{[t]}$ together cost $2E\rho_b$ after E epochs by Proposition 2.9. All T user nodes cost $\frac{1}{T}\sum_{t=1}^{T} 2E\rho_b = 2E\rho_b$ by the *parallel composition theorem* in Proposition 2.10. By the connection of zCDP and (ϵ, δ)-DP in Proposition 2.8, we get $\rho_b = \frac{\epsilon^2}{8E\log(1/\delta)}$, which concludes our proof. □

4.3 Global Variables Update

The server receives T local feature matrix updates, and then updates the global feature matrices according to the same objective function in (5). The gradient for the global feature matrices $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{C}}$ are:

$$\begin{aligned} \widehat{\mathbf{B}} &\leftarrow \widehat{\mathbf{B}} + \eta \sum_{t=1}^{T} \gamma\left({}_{priv}\mathbf{B}^{[t]} - \widehat{\mathbf{B}}\right) \\ \widehat{\mathbf{C}} &\leftarrow \widehat{\mathbf{C}} + \eta \sum_{t=1}^{T} \gamma\left({}_{priv}\mathbf{C}^{[t]} - \widehat{\mathbf{C}}\right). \end{aligned} \tag{19}$$

The update makes the global phenotypes similar to the local phenotypes at the T local sites. The server then sends the global information, $\widehat{\mathbf{B}}, \widehat{\mathbf{C}}$ to each site for the next epoch.

5 EXPERIMENTAL EVALUATION

We evaluate DPFact on three aspects: 1) efficiency based on accuracy and communication cost; 2) utility of the phenotype discovery; and 3) impact of privacy. The evaluation is performed on both real-world datasets and synthetic datasets.

5.1 Dataset

We evaluated DPFact on one synthetic dataset and two real-world datasets, MIMIC-III [17] and the CMS DE-SynPUF[1] dataset. Each of the dataset has different sizes, sparsity (i.e., % of non-zero elements), and skewness in distribution (i.e., some sites have more patients).

[1] https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.html

MIMIC-III.—This is a publicly-available intensive care unit (ICU) dataset collected from 2001 to 2012. We construct 6 local tensors with different sizes representing patients from different ICUs. Each tensor element represents the number of co-occurrence of diagnoses and procedures from the same patient within a 30-day time window. For better interpretability, we adopt the rule in [19] and select 202 procedures ICD-9 codes and 316 diagnoses codes that have the highest frequency. The resulting tensor is 40, 662 patients × 202 procedures × 316 diagnoses with a non-zero ratio of 4.0382×10^{-6}.

CMS.—This is a publicly-available Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF) from 2008 to 2010. We randomly choose 5 samples out of the 20 samples of the outpatient data to construct 5 local tensors with patients, procedures and diagnoses. Different from MIMIC-III, we make each local tensor the same size. There are 82,307 patients with 2,532 procedures and 10,983 diagnoses within a 30-day time window. We apply the same rule in selecting ICD-9 codes. By concatenating the 5 local tensors, we obtain a big tensor with 3.1678×10^{-7} non-zero ratio.

Synthetic Dataset.—We also construct tensors from synthetic data. In order to test different dimensions and sparsities, we construct a tensor of size 5000 × 300 × 800 with a sparsity rate of 10^{-5} and then horizontally partition it into 5 equal parts.

5.2 Baselines

We compare our DPFact framework with two centralized baseline methods and an existing state-of-the-art federated tensor factorization method as described below.

CP-ALS: A widely used, centralized model that solves tensor decomposition using an alternating least squares approach. Data from multiple sources are combined to construct the global tensor.

SGD: A centralized method that solves the tensor decomposition use the stochastic gradient descent-based approach. This is equivalent to DPFact with a single site and no regularization ($T = 1,\ \gamma = 0,\ \mu = 0$). We consider this a counterpart to the CP-ALS method.

TRIP [20]: A federated tensor factorization framework that enforces a shared global model and does not offer any differential privacy guarantee. TRIP utilizes the consensus ADMM approach to decompose the problem into local subproblems.

5.3 Implementation Details

DPFact is implemented in MatlabR2018b with the Tensor Toolbox Version 2.6 [1] for tensor computing and the Parallel Computing Toolbox of Matlab. The experiments were conducted on m5.4xlarge instances of AWS EC2 with 8 workers. For prediction task, we build the logistic regression model with Scikit-learn library of Python 2.7. For reproducibility purpose, we made our code publicly available[2].

[2] https://github.com/jma78/DPFact.

5.4 Parameter Configuration

Hyper-parameter settings include quadratic penalty parameter γ, $l_{2,1}$ regularization term μ, learning rate η, and the input per-epoch, per-factor matrix privacy budget ρ. The rank R is set to 50 to allow some site-specific phenotypes to be captured.

5.4.1 Quadratic penalty parameter γ.—The quadratic penalty term can be viewed as an elastic force between the local factor matrices and the global factor matrices. Smaller γ allows more exploration of the local factors but will result in slower convergence. To balance the trade-off between convergence and stability, we choose $\gamma = 5$ after grid search through $\gamma = \{2, 5, 8, 10\}$.

5.4.2 $l_{2,1}$-regularization term μ.—We evaluate the performance of DPFact with different μ for different ICU types as they differ in the *Lipschitz* constants. Smaller μ has minimal effect on the column sparsity, as there are no columns that are set to 0, while higher μ will "turn off" a large portion of the factors and prevent DPFact from generating useful phenotypes. Based on figure 4 in [23], we choose $\mu = \{1, 1.8, 3.2, 1.8, 1.5, 0.6\}$ for TSICU, SICU, MICU, CSRU, CCU, NICU respectively for MIMIC-III to maintain noticeable differences in the column magnitude and the flexibility to have at least one unshared column (see Appendix in [23] for details). Similarly, we choose $\mu = 2$ equally for each site for CMS and $\mu = 0.5$ equally for each site for the synthetic dataset.

5.4.3 Learning rate η.—The learning rate η must be the same for local sites and the parameter server. The optimal η was found after grid searching in the range $[10^{-5}, 10^{-1}]$. We choose 10^{-2}, 10^{-3}, and 10^{-2} for MIMIC-III, CMS, and synthetic data respectively.

5.4.4 Privacy budget ρ.—We choose the per-epoch privacy budget under the zCDP definition for each factor matrix as $\rho = 10^{-3}$ for MIMIC-III, CMS, and synthetic dataset. By Theorem 4.1, the total privacy guarantee is $(1.2, 10^{-4})$, $(1.9, 10^{-4})$, and $(1.7, 10^{-4})$ under the (ϵ, δ)-DP definition for MIMIC-III, CMS, and synthetic dataset respectively when DPFact converges (we choose δ to be 10^{-4}).

5.4.5 Number of sites T.—To gain more knowledge on how communication cost would be reduced regarding the number of sites, we evaluate the communication cost when the number of sites (T) are increased. To simulate a larger number of sites, we randomly partition the global observed tensor into 1, 5, and 10 sites for the three datasets. Table 2 shows that the communication cost of DPFact scales proportionally with the number of sites.

5.5 Efficiency

5.5.1 Accuracy.—Accuracy is evaluated using the root mean square error (RMSE) between the global observed tensor and a horizontal concatenation of each factorized local tensor. Figure 2 illustrates the RMSE as a function of the number of epochs. We observe that DPFact converges to a smaller RMSE than CP-ALS and TRIP. SGD achieves the lowest RMSE as DPFact suffers some utility loss by sharing differentially private intermediary results.

5.5.2 Communication Cost.—The communication cost is measured based on the total number of communicated bytes divided by the data transfer rate (assumed as 15 MB/second). As CP-ALS and SGD are both centralized models, only TRIP and DPFact are compared.

Table 3 summarizes the communication cost on all the datasets. DPFact reduces the cost by 46.6%, 37.7%, and 20.7% on MIMIC-III, CMS, and synthetic data, respectively. This is achieved by allowing more local exploration at each site (multiple passes of the data) and transmitting fewer auxiliary variables. Moreover, the reduced communication cost does not result in higher RMSE (see Figure 2).

5.6 Utility

The utility of DPFact is measured by the predictive power of the discovered phenotypes. A logistic regression model is fit using the patients' membership values (i.e., $\mathbf{A}_{i:}^{[t]}, \widehat{\mathbf{A}_{i:}}$ of size 1 × R) as features to predict in-hospital mortality. We use a 60–40 train-test split and evaluated the model using area under the receiver operating characteristic curve (AUC).

5.6.1 Global Patterns.—Table 4 shows the AUC for DPFact, CP-ALS (centralized), and TRIP (distributed) as a function of the rank (R). From the results, we observe that DPFact outperforms both baseline methods for achieving the highest AUC. This suggests that DPFact captures similar global phenotypes as the other two methods. We note that DPFact has a slightly lower AUC than CP-ALS for a rank of 10, as the $l_{2,1}$-regularization effect is not prominent.

5.6.2 Site-Specific Patterns.—Besides achieving the highest predictive performance, DPFact also can be used to discover site-specific patterns. As an example, we focus on the neonatal ICU (NICU) which has a drastically different population than the other 5 ICUs. The ability to capture NICU-specific phenotypes can be seen in the AUC comparison with TRIP (Figure 3(a)). DPFact consistently achieves higher AUC for NICU patients. The importance of the $l_{2,1}$-regularization term is also illustrated in Table 4. DPFact with the $l_{2,1}$-regularization is more stable and achieves higher AUC compared without the regularization term ($\mu = 0$).

Table 5 illustrates the top 5 phenotypes with respect to the magnitude of the logistic regression coefficient (mortality risk related to the phenotype) for NICU. The phenotypes are named according to the non-zero procedures and diagnoses. A high λ and prevalence means this phenotype is common. From the results, we observe that heart disease, respiration failure, and pneumonia are more common but less associated with mortality risk (negative coefficient). However, acute kidney injury (AKI) and anemia are less prevalent and highly associated with death. In particular, AKI has the highest risk of in-hospital death, which is consistent with other reported results [33]. Table 6(a) shows an NICU-specific phenotype, which differs slightly from the corresponding global phenotype showing in table 6(b).

5.7 Privacy

We investigated the impact of differential privacy by comparing DPFact with its non-private version. The main difference is that non-private DPFact does not perturb the local feature factor matrices that are transferred to the server. We use the factor match score (FMS) [5] to compare the similarity between the phenotype discovered using DPFact and non-private DPFact. FMS is defined as:

$$score(\overline{\mathcal{X}}) = \frac{1}{R}\sum_{r}\left(1 - \frac{\left|\xi_r - \bar{\xi}_r\right|}{max\left\{\xi_r, \bar{\xi}_r\right\}}\right)\prod_{\mathbf{x}=\mathbf{a},\mathbf{b},\mathbf{c}}\frac{\mathbf{x}_r^T\bar{\mathbf{x}}_r}{\|\mathbf{x}_r\|\|\bar{\mathbf{x}}_r\|},$$

$$\xi_r = \prod_{\mathbf{x}=\mathbf{a},\mathbf{b},\mathbf{c}}\|\mathbf{x}_r\|, \bar{\xi}_r = \prod_{\mathbf{x}=\mathbf{a},\mathbf{b},\mathbf{c}}\|\bar{\mathbf{x}}_r\|$$

where $\overline{\mathcal{X}} = \|\overline{\mathbf{A}}, \overline{\mathbf{B}}, \overline{\mathbf{C}}\|$ is the estimated factors and $\mathcal{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ are the true factors. $\mathbf{x}_r$ are the r^{th} column of factor matrices.

We treat the non-private version DPFact factors as the benchmark for DPFact factors. Figure 3(b) shows how the FMS changes with an increase of the privacy budget. As the privacy budget becomes larger, the FMS increases accordingly and will gradually approximate 1, which means the discovered phenotypes between the two methods are equivalent. This result indicates that when a stricter privacy constraint is enforced, it may negatively impact the quality of the phenotypes. Thus, there is a practical need to balance the trade-off between privacy and phenotype quality.

Table 6 presents a comparison between the top 1 (highest factor weight λ_r) phenotype DPFact-derived phenotype and the closest phenotype derived by its non-private version. We observe that DPFact contains several additional noisy procedure and diagnosis elements than the non-private version DPFact. These extra elements are the results of adding noise to the feature factor matrices. This is also supported in Table 4 as the non-private DPFact has better predictive performance than DPFact. Thus, the output perturbation process may interfere with the interpretability and meaningfulness of the derived phenotypes. However, there is still some utility from the DPFact-derived phenotypes as experts can still distinguish this phenotype to be a heart failure phenotype. Therefore, DPFact still retains the ability to perform phenotype discovery.

6 RELATED WORK

6.1 Tensor Factorization

Tensor analysis is an active research topic and has been widely applied to healthcare data [15, 20, 28], especially for computational phenotyping. Moreover, several algorithms have been developed to scale tensor factorization. GigaTensor [18] used MapReduce for large scale CP tensor decomposition that exploits the sparseness of the real world tensors. DFacTo [6] improves GigaTensor by exploring properties related to the Khatri-Rao Product and achieves faster computation time and better scalability. FlexiFaCT [3] is a scalable MapReduce algorithm for coupled matrix-tensor decomposition using stochastic gradient

descent (SGD). ADMM has also been proved to be an efficient algorithm for distributed tensor factorization [20]. However, the above proposed algorithms have the same potential limitation: the distributed data exhibits the same pattern at different local sites. That means each local tensor can be treated as a random sample from the global tensor. Thus, the algorithms are unable to model the scenario where the distribution pattern may be different at each sites. This is common in healthcare as different units (or clinics and hospitals) will have different patient populations, and may not exhibit all the computational phenotypes.

6.2 Differentially Private Factorization

Differential privacy is widely applied to machine learning areas, especially matrix/tensor factorization, as well as on different distributed optimization frameworks and deep learning problems. Regarding tensor decomposition, there are four ways to enforce differential privacy: input perturbation, output perturbation, objective perturbation and the gradient perturbation. [16] proposed an objective perturbation method for matrix factorization in recommendation systems. [22] proposed a new idea that sampling from the posterior distribution of a Bayesian model can sufficiently guarantee differential privacy. [2] compared the four different perturbation method on matrix factorization and drew the conclusion that input perturbation is the most efficient method that has the least privacy loss on recommendation systems. [27] is the first proposed differentially private tensor decomposition work. It proposed a noise calibrated tensor power method. Our goal in this paper is to develop a distributed framework where data is stored at different sources, and try to preserve the privacy during knowledge transfer. Nevertheless, these works are based on a centralized framework. [20] developed a federated tensor factorization framework, but it simply preserves privacy by avoiding direct patient information sharing, rather than by applying rigorous differential privacy techniques.

7 CONCLUSION

DPFact is a distributed large-scale tensor decomposition method that enforces differential privacy. It is well-suited for computational phenotype from multiple sites as well as other collaborative healthcare analysis with multi-way data. DPFact allows data to be stored at different sites without requiring a single centralized location to perform the computation. Moreover, our model recognizes that the learned global latent factors need not be present at all sites, allowing the discovery of both shared and site-specific computational phenotypes. Furthermore, by adopting a communication-efficient EASGD algorithm, DPFact greatly reduces the communication overhead. DPFact also successfully tackles the privacy issue under the distributed setting with limited privacy loss by the application of zCDP and the parallel composition theorem. Experiments on real-world and synthetic datasets demonstrate that our model outperforms other state-of-the-art methods in terms of communication cost, accuracy, and phenotype discovery ability. Future work will focus on the asynchronization of the collaborative tensor factorization framework to further optimize the computation efficiency.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation, award IIS-#1838200, National Institute of Health (NIH) under award number R01GM114612, R01GM118609, and U01TR002062, and the National Institute of Health Georgia CTSA UL1TR002378. Dr. Xiaoqian Jiang is CPRIT Scholar in Cancer Research, and he was supported in part by the CPRIT RR180012, UT Stars award.

REFERENCES

[1]. Bader Brett W., Kolda Tamara G., et al. 2017 MATLAB Tensor Toolbox Version 3.0-dev. Available online. (Aug. 2017). https://gitlab.com/tensors/tensor_toolbox

[2]. Berlioz Arnaud, Friedman Arik, Kaafar Mohamed Ali, Boreli Roksana, and Berkovsky Shlomo. 2015 Applying differential privacy to matrix factorization. In Proceedings of the 9th ACM Conference on Recommender Systems ACM, 107–114.

[3]. Beutel Alex, Talukdar Partha Pratim, Kumar Abhimanu, Faloutsos Christos, Papalexakis Evangelos E, and Xing Eric P. 2014 Flexifact: Scalable flexible factorization of coupled tensors on hadoop. In Proceedings of the 2014 SDM 109–117.

[4]. Bun Mark and Steinke Thomas. 2016 Concentrated differential privacy: Simplifications, extensions, and lower bounds. In Theory of Cryptography Conference Springer, 635–658.

[5]. Chi Eric C and Kolda Tamara G. 2012 On tensors, sparsity, and nonnegative factorizations. SIAM f. Matrix Anal. Appl. 33, 4 (2012), 1272–1299.

[6]. Choi Joon Hee and Vishwanathan S. 2014 DFacTo: Distributed factorization of tensors. In NIPS 1296–1304.

[7]. Combettes Patrick L and Pesquet Jean-Christophe. 2011 Proximal splitting methods in signal processing In Fixed-point algorithms for inverse problems in science and engineering. Springer, 185–212.

[8]. Dwork Cynthia, Roth Aaron, et al. 2014 The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science 9, 3–4 (2014), 211–407.

[9]. Dwork Cynthia and Rothblum Guy N. 2016 Concentrated differential privacy. arXiv preprint arXiv:1603.01887 (2016).

[10]. Dwork Cynthia, Rothblum Guy N, and Vadhan Salil. 2010 Boosting and differential privacy. In 2010 IEEE 51st Annual Symposium on Foundations of Computer Science IEEE, 51–60.

[11]. Fredrikson Matt, Jha Somesh, and Ristenpart Thomas. 2015 Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security ACM, 1322–1333.

[12]. Greenhalgh Trisha, Hinder Susan, Stramer Katja, Bratan Tanja, and Russell Jill. 2010 Adoption, non-adoption, and abandonment of a personal electronic health record: case study of HealthSpace. Bmj 341 (2010), c5814. [PubMed: 21081595]

[13]. Guo Yuhong and Xue Wei. 2013 Probabilistic Multi-Label Classification with Sparse Feature Learning. In IfCAI 1373–1379.

[14]. Hitaj Briland, Ateniese Giuseppe, and Perez-Cruz Fernando. 2017 Deep models under the GAN: information leakage from collaborative deep learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security ACM, 603–618.

[15]. Ho Joyce C, Ghosh Joydeep, and Sun Jimeng. 2014 Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In Proceedings of the 20th ACM SIGKDD ACM, 115–124.

[16]. Hua Jingyu, Xia Chang, and Zhong Sheng. 2015 Differentially Private Matrix Factorization. In IfCAI 1763–1770.

[17]. Johnson Alistair EW, Pollard Tom J, Shen Lu, Li-wei H Lehman, Feng Mengling, Ghassemi Mohammad, Moody Benjamin, Szolovits Peter, Celi Leo Anthony, and Mark Roger G. 2016 MIMIC-III, a freely accessible critical care database. Scientific data 3 (2016), 160035. [PubMed: 27219127]

[18]. Kang U, Papalexakis Evangelos, Harpale Abhay, and Faloutsos Christos. 2012 Gigatensor: scaling tensor analysis up by 100 times-algorithms and discoveries. In Proceedings of the 18th ACM SIGKDD ACM, 316–324.

[19]. Kim Yejin, El-Kareh Robert, Sun Jimeng, Yu Hwanjo, and Jiang Xiaoqian. 2017 Discriminative and distinct phenotyping by constrained tensor factorization. Scientific reports 7, 1 (2017), 1114. [PubMed: 28442772]

[20]. Kim Yejin, Sun Jimeng, Yu Hwanjo, and Jiang Xiaoqian. 2017 Federated tensor factorization for computational phenotyping. In Proceedings of the 23rd ACM SIGKDD ACM, 887–895.

[21]. Liu Jun, Ji Shuiwang, and Ye Jieping. 2009 Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization. In UAI 339–348.

[22]. Liu Ziqi, Wang Yu-Xiang, and Smola Alexander. 2015 Fast differentially private matrix factorization. In Proceedings of the 9th ACM RecSys 171–178.

[23]. Ma Jing, Zhang Qiuchen, Lou Jian, Ho Joyce C, Xiong Li, and Jiang Xiaoqian. 2019 Privacy-Preserving Tensor Factorization for Collaborative Health Data Analysis. arXiv preprint arXiv: 1908.09888 (2019).

[24]. Nie Feiping, Huang Heng, Cai Xiao, and Ding Chris H. 2010 Efficient and robust feature selection via joint ℓ2,1-norms minimization. In NeurIPS 1813–1821.

[25]. Richesson Rachel L, Sun Jimeng, Pathak Jyotishman, Kho Abel N, and Denny Joshua C. 2016 Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. Artificial intelligence in medicine 71 (2016), 57–61. [PubMed: 27506131]

[26]. Shokri Reza, Stronati Marco, Song Congzheng, and Shmatikov Vitaly. 2017 Membership inference attacks against machine learning models. In Security and Privacy (SP), 2017 IEEE Symposium on IEEE, 3–18.

[27]. Wang Yining and Anandkumar Anima. 2016 Online and differentially-private tensor decomposition. In NeurIPS 3531–3539.

[28]. Wang Yichen, Chen Robert, Ghosh Joydeep, Denny Joshua C, Kho Abel, Chen You, Malin Bradley A, and Sun Jimeng. 2015 Rubik: Knowledge guided tensor factorization and completion for health data analytics. In Proceedings of the 21th ACM SIGKDD ACM, 1265–1274.

[29]. Wei Wei-Qi and Denny Joshua C. 2015 Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome medicine 7, 1 (2015), 41. [PubMed: 25937834]

[30]. Wu Xi, Li Fengan, Kumar Arun, Chaudhuri Kamalika, Jha Somesh, and Naughton Jeffrey. 2017 Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In Proceedings of the 2017 ACM International Conference on Management of Data ACM, 1307–1322.

[31]. Xu Xiao, Li Shu-Xia, Lin Haiqun, Normand SL, Lagu Tara, Desai Nihar, Duan Michael, Kroch Eugene A, and Krumholz Harlan M. 2016 Hospital Phenotypes in the Management of Patients Admitted for Acute Myocardial Infarction. Medical care 54, 10 (2016), 929–936. [PubMed: 27261637]

[32]. Yang Yi, Shen Heng Tao, Ma Zhigang, Huang Zi, and Zhou Xiaofang. 2011 ℓ2,1-norm regularized discriminative feature selection for unsupervised learning. In IfCAI, Vol. 22 1589.

[33]. Youssef Doaa, Abd-Elrahman Hadeel, Shehab Mohamed M, Abd-Elrheem Mohamed, et al. 2015 Incidence of acute kidney injury in the neonatal intensive care unit. Saudi journal of kidney diseases and transplantation 26, 1 (2015), 67. [PubMed: 25579718]

[34]. Yu Lei, Liu Ling, Pu Calton, Gursoy Mehmet Emre, and Truex Stacey. 2019 Differentially Private Model Publishing for Deep Learning. arXiv preprint arXiv:1904.02200 (2019).

[35]. Zhang Sixin, Choromanska Anna E, and LeCun Yann. 2015 Deep learning with elastic averaging SGD. In NeurIPS 685–693.

CCS CONCEPTS

• **Security and privacy** → **Privacy-preserving protocols;** • **Computing methodologies** → **Factorization methods;** • **Applied computing** → *Health informatics*

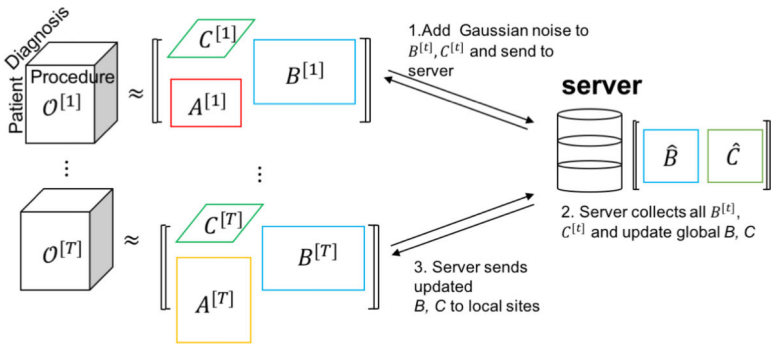
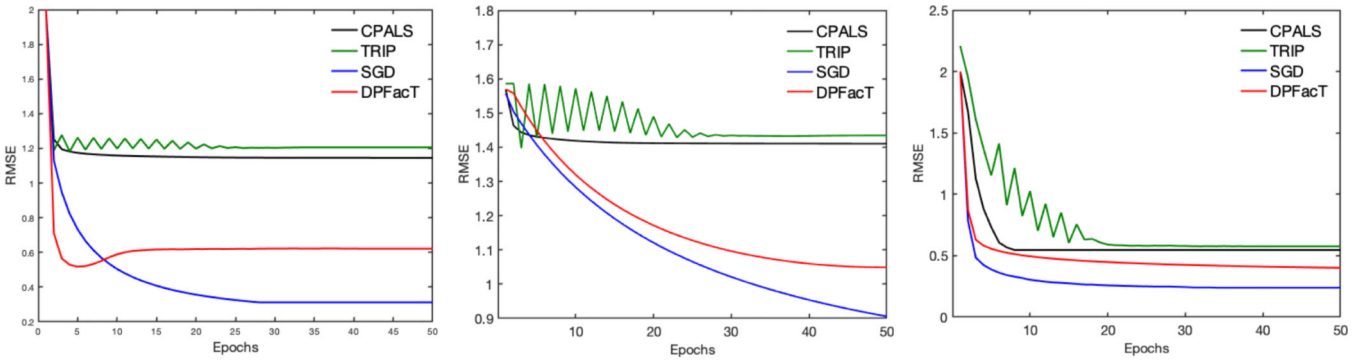


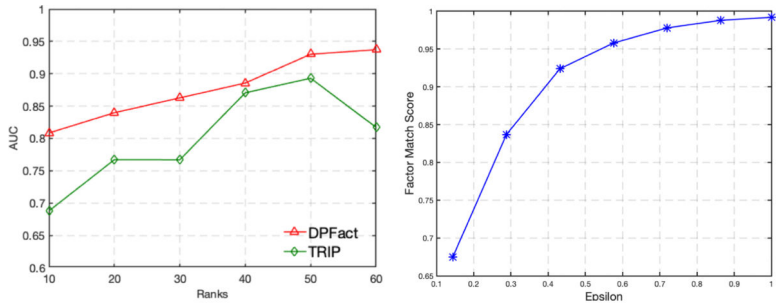
Figure 1:
Algorithm Overview



(a) MIMIC-III: ICU (b) CMS (c) Synthetic

Figure 2:

Average RMSE on (a) MIMIC-III, (b) CMS, (c) Synthetic datasets using 5 random initializations.



(a) AUC comparison (b) Factor Match Score (FMS)

Figure 3:

(a) Predictive performance (AUC) comparison for NICU between (1) TRIP, (2) DPFact. (b) Factor Match Score (FMS) under different privacy budget (ϵ).

Table 1:

Symbols and Notations

Symbols	Descriptions
$\otimes$	Kronecker product
$\odot$	Khatri-Rao product
$\circ$	Outer Product
$*$	Element-wise Product
N	Number of modes
T	Number of local sites
R	Number of ranks
$X_{(n)}$	n-mode matricization of tensor $\mathcal{O}$
$\mathcal{X}$, **X, x**	Tensor, matrix, vector
$\widehat{\mathbf{B}}, \widehat{\mathbf{C}}$	Global factor matrices
$\mathbf{A}^{[t]}$, $\mathbf{B}^{[t]}$, $\mathbf{C}^{[t]}$	Local factor matrices at the t-th site
$\mathcal{X}^{[t]}$	Local tensor at the t-th site
$\mathbf{x}_{i:}$, $\mathbf{x}_{:r}$	Row vector, Column vector

Table 2:

Communication cost of DPFact for different number of sites (Seconds)

# of Sites	MIMIC-III	CMS	Synthetic
1	18.73	22.89	1.55
5	93.62	114.42	7.75
10	189.83	228.83	15.50

Table 3:

Communication Cost of DPFact and TRIP (Seconds)

Algorithm	MIMIC-III	CMS	Synthetic
TRIP	175.26	183.72	9.77
DPFact	93.62	114.42	7.75

Table 4:

Predictive performance (AUC) comparison for (1) CP-ALS, (2) TRIP, (3) DPFact, (4) DPFact without $l_{2,1}$-norm (w/o $l_{2,1}$), (5) non-private DPFact (w/o DP).

Rank	CP-ALS	TRIP	DPFact		
			DPFact	w/o $l_{2,1}$	w/o DP
10	0.7516	0.7130	0.7319	0.5189	0.7401
20	0.7573	0.7596	0.7751	0.6886	0.7763
30	0.7488	0.7644	0.7679	0.6977	0.7705
40	0.7603	0.7574	0.7737	0.7137	0.7756
50	0.7643	0.7633	**0.7759**	0.7212	*0.7790*
60	0.7648	0.7588	0.7758	*0.7312*	0.7763

Table 5:

Top 5 representative phenotypes from NICU based on the factor weights, $\lambda_r = \|\mathbf{A}_{:r}\|_F \|\mathbf{B}_{:r}\|_F \|\mathbf{C}_{:r}\|_F$. Prevalence is the proportion of patients who have non-zero membership to the phenotype.

Phenotypes	Coef	p-value	λ	Prevalence
25: Congenital heart de-fect	−2.1865	0.005	198	34.32
29: Anemia	3.5047	<0.001	77	13.22
30: Acute kidney injury	**5.8806**	<0.001	68	23.38
34: Pneumonia	−5.1050	<0.001	37	37.58
35: Respiratory failure	−0.9141	<0.001	85	24.40

Table 6:

Example of the representative phenotypes. (a)NICU-specific phenotype of Congenital heart defect; (b) and (c) are the globally shared phenotype of Heart failure, showing the difference of DPFact and non-private DPFact.

(a) NICU-specific Phenotypes discovered by DPFact

Procedures	Diagnoses
Cardiac catheterization	Ventricular fibrillation
Insertion of non-drug-eluting coronary artery stent(s)	Unspecified congenital anomaly of heart
Prophylactic administration of vaccine against other disease	Benign essential hypertension

(b) Globally shared phenotype discovered by DPFact

Procedures	Diagnoses
Attachment of pedicle or flap graft	Rheumatic heart failure
Right heart cardiac catheterization	Ventricular fibrillation
Procedure on two vessels	Benign essential hypertension
Other endovascular procedures on other vessels Insertion of non-drug-eluting coronary artery stent(s)	Paroxysmal ventricular tachycardia Nephritis and nephropathy

(c) Globally shared phenotype discovered by non-private DPFact

Procedures	Diagnoses
Right heart cardiac catheterization	Hypopotassemia
Attachment of pedicle or flap graft	Rheumatic heart failure
Excision or destruction of other lesion or tissue of heart, open approach	Benign essential hypertension Paroxysmal ventricular tachycardia Systolic heart failure