Robust Irregular Tensor Factorization and Completion for Temporal Health Data Analysis

Yifei Ren
yifei.ren2@emory.edu
Emory University

Jian Lou*
jian.lou@emory.edu
Emory University

Li Xiong
lxiong@emory.edu
Emory University

Joyce C. Ho
joyce.c.ho@emory.edu
Emory University

ABSTRACT

Electronic health records (EHR) are often generated and collected across a large number of patients featuring distinctive medical conditions and clinical progress over a long period of time, which results in unaligned records along the time dimension. EHR is also prone to missing and erroneous data due to various practical reasons. Recently, PARAFAC2 has been re-popularized for successfully extracting meaningful medical concepts (phenotypes) from such temporal EHR by irregular tensor factorization. Despite recent advances, existing PARAFAC2 methods are unable to robustly handle erroneousness and missing data which are prevalent in clinical practice. We propose **REPAIR**, a **R**obust t**E**mporal **PA**RAFAC2 method for **IR**regular tensor factorization and completion method, to complete an irregular tensor and extract phenotypes in the presence of missing and erroneous values. To achieve this, REPAIR designs a new effective low-rank regularization function for PARAFAC2 to handle missing and erroneous entries, which has not been explored for irregular tensors before. In addition, the optimization of REPAIR allows it to enjoy the same computational scalability and incorporate a variety of constraints as the state-of-the-art PARAFAC2 method for efficient and meaningful phenotype extraction. We evaluate REPAIR on two real temporal EHR datasets to verify its robustness in tensor factorization against various missing and outlier conditions. Furthermore, we conduct two case studies to demonstrate that REPAIR is able to extract meaningful and useful phenotypes from such corrupted temporal EHR. Our implementation is publicly available[1].

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Computing methodologies** → *Factor analysis*.

KEYWORDS

Tensor Factorization; Electronic Health Records (EHR); PARAFAC2; Irregular Tensor; Computational Phenotyping

ACM Reference Format:
Yifei Ren, Jian Lou, Li Xiong, and Joyce C. Ho. 2020. Robust Irregular Tensor Factorization and Completion for Temporal Health Data Analysis. In *The 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340531.3411982

1 INTRODUCTION

Tensors are a popular algebraic structure for a wide range of applications, due to their exceptional capability to model multidimensional relationships of the data. Among them, regular tensors with aligned dimensions for all modes have been extensively studied, for which various tensor factorization structures are proposed depending on the applications, e.g. Canonical Polyadic (CP) [8, 13, 15], Tucker [35], and tensor singular value decomposition (SVD) [21, 22]. On the contrary, the irregular tensor with unaligned size along one of its modes is under studied [4, 14], despite its prevalence in real-world practice.

As the motivating example considered in this paper, electronic health records (EHR) are datasets collected during clinic practice, which encompasses clinical records of a large number of distinct patients across a long period of time. EHR data do not always directly and reliably map to medical concepts that clinical researchers need or use [20]. Tensor factorization methods have shown great potential in discovering meaningful and interpretable clinical concepts (or phenotypes) from complicated health records [2, 16–18, 32, 37]. The resulting tensor factors are reported as phenotype candidates that automatically reveal patient clusters on specific diagnoses and procedures [9]. Such analysis can be particularly useful for understanding disease subtypes and clinical progressions in different subpopulations for new and rapidly evolving diseases such as the current COVID-19 pandemic. Yet, temporal EHR data poses additional challenges for phenotype analysis due to: 1) the irregularity of the data along the time dimension, and 2) the potentially missing and erroneous entries during the data collection over long period of time. Concretely, the records are unaligned in time from patient to patient because of the varying disease states and progressions, which lead to variable number of clinic encounters and different time gaps between consecutive visits. In addition, they are prone to corruptions due to various reasons during clinical practice, for

*Corresponding Author.
[1]https://github.com/Emory-AIMS/Repair

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

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3411982

example, equipment failure, inexperienced clinical staff, and inaccurate information recording. As a result, most existing tensor factorization frameworks for regular tensors are not applicable or do not work well.

PARAFAC2 [12] is a dedicated multidimensional algebraic framework for modeling irregular tensors, which has a natural factorization structure for handling variable sizes along the unaligned mode. In particular, for temporal EHR, recent advances have improved its computational scalability as well as its factorization interpretability [2]. Despite these improvements, existing PARAFAC2 methods are not robust to missing and erroneous elements in the data, which severely limits its applicability to practical temporal EHR data analysis.

For regular tensor factorization frameworks, robust mechanisms are well developed to handle missing and erroneous data, among which the robust low-rank tensor minimization (RLTM) is one of the most successful approaches [1, 10, 11, 25–27, 29, 34]. Different low-rank regularization functions are adopted by these methods, which vary according to different types of tensor factorization. However, it is still unknown how to impose low-rank regularization for PARAFAC2 and design an explicit RLTM mechanism to handle missing entries and remove erroneous entries.

To fill this gap, we propose **REPAIR**, a **R**obust t**E**mporal **PA**FAFAC2 method for **IR**regular tensor factorization and completion (c.f. Figure 1), which is the first robust irregular tensor recovery method. Given each patient input data $\mathbf{O}_k$ with erroneous and missing entries, REPAIR performs RLTM to separate out the erroneous entries $\mathbf{E}_k$ from the underlying clean and completed components $\mathbf{X}_k$, and uses the clean tensor to form a common low rank space for PARAFAC2 based candidate phenotype extraction, i.e. $\mathbf{X}_k \approx \mathbf{U}_k S_k \mathbf{V}^\top$. We achieve this by addressing two main challenges: First, specific low-rank regularizations need to be designed for PARAFAC2 to suit its decomposition structure which has not been explored in existing work. Second, the robust factorization needs to incorporate additional constraints such as temporal smoothness, non-negativity and sparsity [2] to obtain more meaningful and accurate phenotypes.

We summarize our contributions below:

(1) We propose a robust PARAFAC2 tensor factorization method for irregular tensors with a new low-rank regularization function to handle potentially missing and erroneous entries in the input tensor. This is the first work that explicitly handles missing and erroneous data for irregular tensor factorization.
(2) We design an efficient two-phase optimization to simultaneously: 1) learn and complete the clean underlying tensor by decomposing the original tensor $\{\mathbf{O}_k\}$ into the underlying low-rank tensor $\{\mathbf{X}_k\}$ and the sparse error tensor $\{\mathbf{E}_k\}$ (Fig. 1 blue box); and 2) extract phenotypes by factorizing the clean tensor $\mathbf{X}_k = \mathbf{U}_k \mathbf{S}_k \mathbf{V}^\top$ (Fig. 1 red box). The phenotype extraction phase incorporates many practical constraints for improving interpretability of the extracted phenotypes, including temporal smoothness, non-negativity and sparsity.
(3) We evaluate **REPAIR** on two real-world temporal EHR datasets with a set of experiments, which verify the improved recovery and factorization robustness against missing and erroneous values. Through two case studies: identification of higher-risk patient subgroups, and in-hospital mortality prediction, we further demonstrate the superior utility of the factorization outputs of REPAIR to facilitate downstream temporal EHR data analysis.

Table 1: Symbols and notations used in this paper

Symbol	Definition
$\mathbf{a}, \mathbf{A}, \boldsymbol{\mathcal{A}}$	Vector, Matrix, Tensor
$\mathbf{A}_k$	k-th frontal slice of $\boldsymbol{\mathcal{A}}$
$\boldsymbol{\mathcal{A}}_{(n)}$	Mode-n matricization of $\boldsymbol{\mathcal{A}}$
$\|\cdot\|_1$	ℓ_1-norm
$\|\cdot\|_F$	Frobenius norm
$\|\cdot\|_*$	Nuclear norm
$*$	Hadamard (element-wise) multiplication
$\odot$	Khatri Rao product
$\circ$	Outer product
$\langle\cdot,\cdot\rangle$	Inner product

2 BACKGROUND

In this section, we define the notations and present background on robust low-rank tensor minimization followed by PARAFAC2 and its application for temporal EHR phenotyping. Table 1 summarizes commonly used notations.

For temporal EHR, let the observed tensor be $\mathcal{O} = \{\mathbf{O}_k\} \in \{\mathbb{R}^{I_k \times J}\}$ (c.f. leftmost tensor in Figure 1) with 3 modes, where each frontal slice $\mathbf{O}_k$ represents patient k's record of J types of diagnosis, treatments or lab test results (along mode 2), across I_k clinical encounters (along mode 1) varying from patient to patient. The aim of temporal EHR phenotyping is to discover medical concepts by making use of all K frontal slices, i.e. the information of all K patients, and discerning as much inter-relationship across different patients (i.e. across frontal slice) as possible.

2.1 Robust Low-rank Tensor Factorization and Completion

For regular tensors (i.e. assuming $\{\mathbf{O}_k\}$ are aligned in all dimensions), the robust low-rank tensor minimization (RLTM) is one of the most successful approaches to handle incomplete and corrupted input tensors. For such a regular tensor $\mathcal{O}$, RLTM separates it into an underlying clean and completed tensor $\mathcal{X}$ and an error tensor $\mathcal{E}$. In practice, the clean part is often low-rank while the erroneous part is sparse. Thus, RLTM imposes a low-rank regularization function $\|\cdot\|_{lr}$ and a sparsity regularization function $\|\cdot\|_1$ on $\mathcal{X}$ and $\mathcal{E}$, correspondingly:

$$\underset{\mathcal{X},\mathcal{E}}{\operatorname{argmin}} \|\mathcal{X}\|_{lr} + \rho_0 \|\mathcal{E}\|_1, \; s.t. \; \mathcal{P}_\Omega(\mathcal{O}) = \mathcal{P}_\Omega(\mathcal{X} + \mathcal{E}), \qquad (1)$$

where Ω is the index set of non-missing entries and $\mathcal{P}_\Omega$ keeps entries in Ω and zeros out others (i.e., missing entries), ρ_0 is a balancing constant. RLTM is a multidimensional extension to the robust low-rank matrix minimization [7], but it is intrinsically more difficult. The main challenge lies in introducing a proper low-rank definition and designing an effective and efficient low-rank regularization. Unlike a low-rank matrix, the low-rank definition

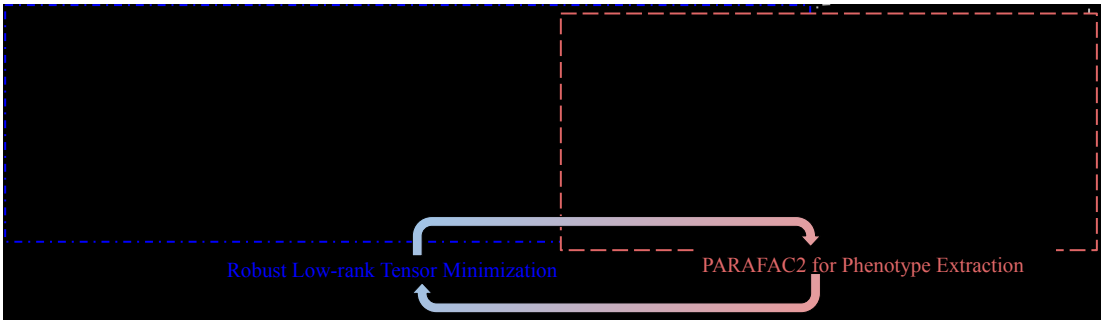


Figure 1: Overview of REPAIR: robust irregular tensor PARAFAC2 factorization for EHR phenotyping on input patients' data $\mathcal{O}$. $\mathbf{O}_k$ contains erroneous and missing entries, which can be decomposed into erroneous $\mathbf{E}_k$ and clean and completed components denoted by $\mathbf{X}_k$. $\mathcal{P}_\Omega(\mathbf{O}_k) = \mathcal{P}_\Omega(\mathbf{E}_k + \mathbf{X}_k)$. The underling clean tensor is decomposed by PARAFAC2 into $\mathbf{X}_k \approx \mathbf{U}_k S_k \mathbf{V}^\top$.

for tensor is not unique and should be adapted according to each tensor decomposition model (e.g., CP, Tucker, tensor SVD).

For example, Tucker model defines the rank of $\mathcal{X}$ based on the matrix rank of its matricization, i.e. the vector $(rank(\mathcal{X}_{(1)}), rank(\mathcal{X}_{(2)}), rank(\mathcal{X}_{(3)}))$. CP decomposes $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ into the sum of R rank-one tensors by $\mathcal{X} = \sum_{r=1}^{R} \mathbf{A}(:, r) \circ \mathbf{B}(:, r) \circ \mathbf{C}(:, r)$, where $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are factorization matrices and the smallest R to achieve such decomposition is defined to be the rank R^* of $\mathcal{X}$ under CP model. It is difficult to accurately estimating R^* for CP (in fact, NP-hard to determine), as well as to deal with matrix rank used by Tucker. More tractable relaxations are then proposed with various low-rank regularization functions [10, 25, 26, 34].

Despite their varieties, the existing low-rank regularization functions are designed for regular tensor factorization models and cannot be applied to an irregular tensor factorization model like PARAFAC2. In fact, they are not even well-defined on irregular tensors and PARAFAC2. Thus, there lacks a tractable and effective low-rank regularization for PARAFAC2 applicable to large-scale irregular tensors.

2.2 PARAFAC2 for Temporal EHR

PARAFAC2 is the state-of-the-art tensor factorization structure for irregular tensors that do not align naturally along one of its modes. The classic PARAFAC2 (c.f. Fig. 1 red box) for irregular tensor $\{\mathbf{X}_k\}$ is formalized below [24]:

Definition 1. *(Classic PARAFAC2 model)*

$$\underset{\{\mathbf{U}_k\},\{\mathbf{S}_k\},\mathbf{V}}{\operatorname{argmin}} \sum_{k=1}^{K} \frac{1}{2} \|\mathbf{X}_k - \mathbf{U}_k \mathbf{S}_k \mathbf{V}^\top\|_F^2,$$

s.t. $\mathbf{U}_k = \mathbf{Q}_k \mathbf{H}, \mathbf{Q}_k^\top \mathbf{Q}_k = \mathbf{I}, \mathbf{S}_k$ *is diagonal, where* $\mathbf{Q}_k \in \mathbb{R}^{I_k \times R}$ *is orthogonal,* $\mathbf{I}_k \in \mathbb{R}^{R \times R}$ *is the identity matrix and* R *is the target rank of the PARAFAC2 decomposition.*

For temporal EHR data, the factorization matrices have the following interpretation:

- ▷ $\mathbf{U}_k \in \mathbb{R}^{I_k \times R}$ contains the temporal evolution for patient k: the r-th column of $\mathbf{U}_k$ indicates the evolution of the r-th phenotype for all I_k clinical visits for patient k.
- ▷ $\mathbf{V} \in \mathbb{R}^{J \times R}$ reflects the phenotypes. Each non-zero entry of V indicates the membership of the corresponding j-th medical feature in the r-th phenotype.
- ▷ $\mathbf{S}_k \in \mathbb{R}^{R \times R}$ is a diagonal matrix with the importance membership of patient k in each one of the R phenotypes. It is often organized into $\mathbf{W} \in \mathbb{R}_+^{R \times K}$ with each row of $\mathbf{W}$ composed by the diagonal of $\mathbf{S}_k$, i.e. $\mathbf{W}(:, k) = \texttt{diag}(\mathbf{S}_k)$.

SPARTan [32] scales PARAFAC2 to large temporal EHR phenotyping by introducing a sparse MTTKRP (abbreviated for Matricized-Tensor-Times-Khatri-Rao-Product) module, which takes advantage of the high input sparsity to reduce the per-iteration cost. Following its efficiency improvement, COPA [2] further introduces various constraints/regularizations to improve the interpretability of the factor matrices for more meaningful pheonotype extraction. For example, COPA introduces the M-spline constraint [33] to $\mathbf{U}_k$ to capture the temporal smoothness, non-negative constraint to $\mathbf{S}_k$ to get positive weight, and sparsity (e.g., ℓ_1 norm regularization) to $\mathbf{V}$ to induce sparse phenotype definitions.

In sum, despite their improvements on computational efficiency and output interpretability, existing PARAFAC2 methods do not explicitly address the problem of extracting meaningful phenotypes from EHR datasets with moderate ratio of missing and error entries, which severely limits them from more robust clinical usage.

3 PROPOSED METHOD

3.1 Low-rank Regularization for PARAFAC2

As mentioned, the effective low-rank regularization has not been studied for irregular tensors. Recent work [38] proposes to recover each of $\{\mathbf{X}_k\}$'s frontal slices matrix by matrix by robust low-rank matrix completion techniques [6, 28]. The drawback of this approach is that it cannot capture the internal structural correlations across frontal slices, i.e. common information among patients, for temporal EHR phenotyping. As can be seen from our experiments, this approach does not provide satisfactory recovery performance. On the contrary, we propose to impose the low-rankness on $\{\mathbf{X}_k\}$ through adding nuclear norm constraints on the internal factorization matrices $\mathbf{H}, \mathbf{V}, \mathbf{W}$, which are shared by all frontal slices thus capable of capturing cross-slice information.

Definition 2. *For irregular tensor*

$$\mathcal{X} = \{\mathbf{X}_k\} \approx \texttt{PARAFAC2}(\{\mathbf{Q}_k\}, \mathbf{H}, \mathbf{V}, \mathbf{W}),$$

the low-rank regularization function is defined as

$$\|\mathcal{X}\|_{lr} := \|\mathbf{H}\|_* + \|\mathbf{V}\|_* + \|\mathbf{W}\|_*. \tag{2}$$

Our low-rank regularization function enjoys the following nice properties: 1) it is natural to the decomposition structure of the PARAFAC2 model; 2) it can effectively recover the underlying clean and completed tensor $\{\mathbf{X}_k\}$ by capturing cross frontal slice information.

3.2 REPAIR: Model

Having defined the low-rank regularization function in Definition 2, we formalize the objective function for the REPAIR model in Definition 3. It applies the RLTM framework (i.e. eq.(1)) to PARAFAC2, which separates the underlying clean and completed tensor $\mathcal{X} = \{\mathbf{X}_k\}$ and the erroneous tensor $\mathcal{E} = \{\mathbf{E}_k\}$ given the missing and corrupted observation tensor $\mathcal{O} = \{\mathbf{O}_k\}$. Meanwhile, REPAIR decomposes $\{\mathbf{X}_k\}$ into PARAFAC2 structure. The tensor recovery of $\mathcal{X}$ is enforced by the linear constraint between $\mathcal{O}$, $\mathcal{X}$, $\mathcal{E}$ in eq (4), low-rank regularization for $\mathcal{X} = \{\mathbf{X}_k\}$ and sparsity constraint for $\mathcal{E}$ in the first row of eq (3). The tensor factorization of $\mathcal{X}$ is enforced by the PARAFAC2 loss for $\mathcal{X}$, the temporal smoothness, nonnegativity, and sparsity constraints in the second row of eq (3) and additional constraints in eq (5).

For EHR phenotype discovery, various constraints should be imposed on the factorization matrices to yield meaningful and high-interpretability phenotypes. The REPAIR model accommodates such interpretability-purposed constraints in eq.(3) including: temporal smoothness for $c_1(\mathbf{H})$, non-negativity for $\{c_2(\mathbf{S}_k)\}$, sparsity for $c_3(\mathbf{V})$.

Definition 3. *(REPAIR objective function)*

$$\underset{\mathbf{Q}_k, \mathbf{H}, \mathbf{S}_k, \mathbf{V}}{\operatorname{argmin}} \sum_{k=1}^{K} \Big(\overbrace{\|\mathbf{X}_k - \mathbf{U}_k\mathbf{S}_k\mathbf{V}^\top\|_F^2}^{\text{PARAFAC2 loss for } \mathcal{X}} + \overbrace{\rho_0\|\mathcal{P}_\Omega(\mathbf{E}_k)\|_1}^{\text{sparsity for } \mathcal{E}} \Big) +$$

$$\overbrace{\rho_1\|\mathbf{H}\|_* + \rho_2\|\mathbf{V}\|_* + \rho_3\|\mathbf{W}\|_*}^{\text{low-rankness for } \mathcal{X}} + \overbrace{c_1(\mathbf{H})}^{\text{smoothness}} + \overbrace{\sum_{k=1}^{K} c_2(\mathbf{S}_k)}^{\text{nonnegativity}} + \overbrace{c_3(\mathbf{V})}^{\text{sparsity}}, \tag{3}$$

$$s.t.\ \text{for } k = 1, ..., K, \quad \overbrace{\mathcal{P}_\Omega(\mathbf{O}_k) = \mathcal{P}_\Omega(\mathbf{X}_k + \mathbf{E}_k)}^{\text{linear constraint between } \mathcal{O}, \mathcal{X}, \mathcal{E}}, \tag{4}$$

$$\underbrace{\mathbf{S}_k = \texttt{diag}(\mathbf{W}(k,:)), \mathbf{S}_k \text{ is diagonal}}_{\text{relation between } \mathcal{S}, \mathbf{W}}, \quad \underbrace{\mathbf{U}_k = \mathbf{Q}_k\mathbf{H},\ \mathbf{Q}_k^\top\mathbf{Q}_k = \mathbf{I}}_{\text{constraints for PARAFAC2 decomposition}} \tag{5}$$

where $\mathbf{H}, \{\mathbf{S}_k\}, \mathbf{I} \in \mathbb{R}^{R\times R}$, $\mathbf{Q}_k \in \mathbb{R}^{I_k \times R}$.

3.3 REPAIR: Optimization

To solve the REPAIR model, a straightforward approach is to introduce auxiliary variables for the low-rank and interpretability regularizations, then solve the problem by multi-block Alternating Direction Method of Multipliers (ADMM) [3]. Inspired by the more flexible Alternating Optimization ADMM (AO-ADMM) [19], we design a two-phase alternative optimization algorithm to accommodate more constraints. The REPAIR optimization proceeds by iterating between the two phases: I) updating the factorization matrices $\{\mathbf{Q}_k\}, \mathbf{H}, \mathbf{V}, \mathbf{W}$; II) separating the $\mathcal{X}$ and $\mathcal{E}$ from $\mathcal{O}$. For I), we factorize the intermediate (inaccurate) recovered tensor $\mathcal{X}$ by solving an approximated PARAFAC2; for II), we follow standard ADMM to convert the linear constraint of eq.(4) by introducing Lagrangian dual variable $\{\Gamma_O^k\}$ to get rid of the constraint as shown in Definition 3. This way, REPAIR can accommodate a variety of constraints for each factorization for better interpretability. Also, the optimizations for each factor are more independent, which makes it easier to deal with.

Definition 4. *The augmented Lagrangian dual objective is,*

$$\begin{aligned}
&\sum_{k=1}^{K} \Big(\|\mathbf{X}_k - \mathbf{Q}_k\mathbf{H}\mathbf{S}_k\mathbf{V}^\top\|_F^2 - \langle \Gamma_O^k, \mathbf{O}_k - \mathbf{X}_k - \mathbf{E}_k \rangle \\
&+ \frac{\eta_O^k}{2}\|\mathbf{O}_k - \mathbf{X}_k - \mathbf{E}_k\|_F^2 + \rho_0\|\mathcal{P}_\Omega(\mathbf{E}_k)\|_1 \Big) \\
&+ \Big(\rho_1\|\mathbf{H}\|_* + \rho_2\|\mathbf{V}\|_* + \rho_3\|\mathbf{W}\|_* \Big) + \Big(c_1(\mathbf{H}) + c_2(\mathbf{W}) + c_3(\mathbf{V}) \Big) \\
&s.t.\ \mathbf{S}_k = \texttt{diag}(\mathbf{W}(k,:)),\ \mathbf{Q}_k^\top\mathbf{Q}_k = \mathbf{I},\ \text{for } k = 1, ..., K.
\end{aligned}$$

Phase I: Approximated PARAFAC2. In the first phase, we update the factorization matrices $\{\mathbf{Q}_k\}, \mathbf{H}, \mathbf{V}, \mathbf{W}$ with $\{\mathbf{X}_k\}$ and $\{\mathbf{E}_k\}$ fixed, which can be intuitively seen as decomposing the latest recovered tensor $\{\mathbf{X}_k\}$ into PARAFAC2. In practice, we observe that it is enough to run PARAFAC2 by one iteration in this phase to achieve the overall convergence, which avoids heavy computation of solving precise PARAFAC2.

Update $\mathbf{Q}_k$*:* To update $\mathbf{Q}^k$, we need Lemma 1 below:

Lemma 1. *The Orthogonal Procrustes problem is:*

$$\mathbf{Q}^\# = \underset{\mathbf{Q}:\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}}{\operatorname{argmin}} \|\mathbf{Q}\mathbf{A} - \mathbf{B}\|_F^2,$$

which has the closed-form solution: $\mathbf{Q}^\# = \mathbf{P}\mathbf{Z}^\top$*, where* $[\mathbf{P}, \Sigma, \mathbf{Z}] = \texttt{svd}(\mathbf{B}\mathbf{A}^\top)$ *and* $\texttt{svd}(\cdot)$ *is singular value decomposition.*

When applied to the update of $\mathbf{Q}_k$, with other factors fixed, we have

$$\mathbf{Q}_k = \underset{\mathbf{Q}_k:\mathbf{Q}_k^\top\mathbf{Q}_k=\mathbf{I}}{\operatorname{argmin}} \|\mathbf{X}_k - \mathbf{Q}_k\mathbf{H}\mathbf{S}_k\mathbf{V}^\top\|_F^2. \tag{6}$$

Let $\mathbf{B} = \mathbf{X}_k$ and $\mathbf{A} = \mathbf{H}\mathbf{S}_k\mathbf{V}^\top$ and by Lemma 1:

$$\mathbf{Q}_k = \mathbf{P}_k\mathbf{Z}_k^\top, \ \text{where}[\mathbf{P}_k, \Sigma, \mathbf{Z}_k] = \texttt{svd}(\mathbf{X}_k\mathbf{V}\mathbf{S}_k\mathbf{H}^\top). \tag{7}$$

Update $\mathbf{H}$*:* After obtaining $\{\mathbf{Q}_k\}$, we denote $\mathbf{Y}_k = \mathbf{Q}_k^\top\mathbf{X}_k$, for $k = 1, ..., K$, and let $\mathcal{Y}$ be the tensor with $\mathbf{Y}_k$ being its frontal slice. We then update $\mathbf{H}, \mathbf{V}, \mathbf{W}$ alternatively by solving three constrained least squares sub-problems. Due to the symmetry of the three sub-problems, we elaborate the update for $\mathbf{H}$ as an example.

$$\mathbf{H} = \underset{\mathbf{H}}{\operatorname{argmin}} \|\mathcal{Y}_{(1)} - \mathbf{H}(\mathbf{V} \odot \mathbf{W})^\top\|_F^2 + \rho_1\|\mathbf{H}\|_* + c_1(\mathbf{H}).$$

Table 2: Additional Symbols for REPAIR Optimization

Symbol	Definition
$\rho_0, \rho_1, \rho_2, \rho_3$	Balancing hyper-parameters
$\mathbf{H}^l, \mathbf{V}^l, \mathbf{W}^l$	Auxiliary variable for low-rank constr.
$\Gamma_H^l, \Gamma_W^l, \Gamma_V^l$	Lagrangian dual for low-rank constr.
$\eta_H^l, \eta_W^l, \eta_V^l$	Lagrangian constant for low-rank constr.
$\mathbf{H}^c, \mathbf{V}^c, \mathbf{W}^c$	Auxiliary variable for interpretability constr.
$\Gamma_H^c, \Gamma_W^c, \Gamma_V^c$	Lagrangian dual for interpretability constr.
$\eta_H^c, \eta_W^c, \eta_V^c$	Lagrangian constant for interpretability constr.

We introduce two auxiliary variables $\mathbf{H}^l$ and $\mathbf{H}^c$ to separate the low-rank and interpretability constraints:

$$\underset{\mathbf{H},\mathbf{H}^l,\mathbf{H}^c}{\operatorname{argmin}} \|\mathcal{Y}_{(1)} - \mathbf{H}(\mathbf{V}\odot\mathbf{W})^\top\|_F^2 + \rho_1\|\mathbf{H}^l\|_* + c_1(\mathbf{H}^c),$$
$$s.t.\ \mathbf{H}^l = \mathbf{H},\ \mathbf{H}^c = \mathbf{H}.$$

The above can be solved by ADMM after introducing the Lagrangian dual variables Γ_H^l, Γ_H^c and constants η_H^l, η_H^c, correspondingly:

$$\begin{aligned}\underset{\mathbf{H},\mathbf{H}^l,\mathbf{H}^c}{\operatorname{argmin}} \|\mathcal{Y}_{(1)} - \mathbf{H}(\mathbf{V}\odot\mathbf{W})^\top\|_F^2 + \rho_1\|\mathbf{H}^l\|_* + c_1(\mathbf{H}^c)\\ - \langle\Gamma_H^l, \mathbf{H}-\mathbf{H}^l\rangle + \frac{\eta_H^l}{2}\|\mathbf{H}-\mathbf{H}^l\|_F^2\\ - \langle\Gamma_H^c, \mathbf{H}-\mathbf{H}^c\rangle + \frac{\eta_H^c}{2}\|\mathbf{H}-\mathbf{H}^c\|_F^2.\end{aligned}$$

To solve it by ADMM, we have the following update sequence for $\mathbf{H}, \mathbf{H}^l, \mathbf{H}^c$ and dual Γ_H^l, Γ_H^c:

$$\begin{aligned}\mathbf{H} = &\left(\mathcal{Y}_{(1)}(\mathbf{V}\odot\mathbf{W}) + \Gamma_H^l + \Gamma_H^c + \eta_H^l\mathbf{H}^l + \eta_H^c\mathbf{H}^c\right)\\ &\cdot\left((\mathbf{V}^\top\mathbf{V}) * (\mathbf{W}^\top\mathbf{W}) + (\eta_H^l + \eta_H^c)\mathbf{I}\right)^\dagger,\end{aligned} \tag{8}$$

where $\odot$ is the Khatri Rao product, $*$ is the Hadamard product and $\dagger$ is the pseudo-inverse.

$$\mathbf{H}^l = \underset{\mathbf{H}^l}{\operatorname{argmin}} \frac{\eta_H^l}{2}\|\mathbf{H}^l - \mathbf{H}\|_F^2 - \langle\Gamma_H^l, \mathbf{H}^l - \mathbf{H}\rangle + \rho_1\|\mathbf{H}^l\|_*,$$

which has the proximal operator [30] with respect to the nuclear norm $\|\cdot\|_*$, a.k.a. singular value thresholding [6], as its closed-form solution:

$$\mathbf{H}^l = \mathtt{prox}_{\frac{\rho_1}{\eta_H^l}\|\cdot\|_*}(\mathbf{H} + \frac{\Gamma_H^l}{\eta_H^l}) = \mathbf{P}\mathtt{Diag}(\max\{0, \boldsymbol{\sigma} - \frac{\rho_1}{\eta_H^l}\})\mathbf{Z}^\top, \tag{9}$$

where $[\mathbf{P}, \mathtt{Diag}(\boldsymbol{\sigma}), \mathbf{Z}] = \mathtt{svd}(\mathbf{H} + \frac{\Gamma_H^l}{\eta_H^l})$.

$$\mathbf{H}^c = \underset{\mathbf{H}^c}{\operatorname{argmin}} \frac{\eta_H^c}{2}\|\mathbf{H}^c - \mathbf{H}\|_F^2 - \langle\Gamma_H^c, \mathbf{H}^c - \mathbf{H}\rangle + c_1(\mathbf{H}^c),$$

which has the proximal operator with respect to the constraint function $c_1(\cdot)$ as its closed-form solution:

$$\mathbf{H}^c = \mathtt{prox}_{\frac{1}{\eta_H^c}c_1}(\mathbf{H} + \frac{\Gamma_H^c}{\eta_H^c}). \tag{10}$$

The Lagrangian dual variables are update as follows:

$$\Gamma_H^c = \Gamma_H^c - \eta_H^c(\mathbf{H} - \mathbf{H}^c); \tag{11}$$
$$\Gamma_H^l = \Gamma_H^l - \eta_H^l(\mathbf{H} - \mathbf{H}^l). \tag{12}$$

Update $\mathbf{V}, \mathbf{W}$*:* The update for $\mathbf{V}$ (along with $\mathbf{V}^l, \mathbf{V}^c$) and $\mathbf{W}$ (along with $\mathbf{W}^l, \mathbf{W}^c$) are similar to $\mathbf{H}$:

$$\begin{aligned}\mathbf{V} = &\left(\mathcal{Y}_{(2)}(\mathbf{H}\odot\mathbf{W}) + \Gamma_V^l + \Gamma_V^c + \eta_V^l\mathbf{V}^l + \eta_V^c\mathbf{V}^c\right)\\ &\cdot\left((\mathbf{H}^\top\mathbf{H}) * (\mathbf{W}^\top\mathbf{W}) + (\eta_V^l + \eta_V^c)\mathbf{I}\right)^\dagger;\\ \mathbf{V}^l = &\ \mathtt{prox}_{\frac{\rho_3}{\eta_V^l}\|\cdot\|_*}(\mathbf{V} + \frac{\Gamma_V^l}{\eta_V^l}); \mathbf{V}^c = \mathtt{prox}_{\frac{1}{\eta_V^c}c_3}(\mathbf{V} + \frac{\Gamma_V^c}{\eta_V^c});\\ \Gamma_V^c = &\ \Gamma_V^c - \eta_V^c(\mathbf{V} - \mathbf{V}^c); \Gamma_V^l = \Gamma_V^l - \eta_V^l(\mathbf{V} - \mathbf{V}^l).\end{aligned} \tag{13}$$

$$\begin{aligned}\mathbf{W} = &\left(\mathcal{Y}_{(3)}(\mathbf{V}\odot\mathbf{H}) + \Gamma_W^l + \Gamma_W^c + \eta_W^l\mathbf{W}^l + \eta_W^c\mathbf{W}^c\right)\\ &\cdot\left((\mathbf{V}^\top\mathbf{V}) * (\mathbf{H}^\top\mathbf{H}) + (\eta_W^l + \eta_W^c)\mathbf{I}\right)^\dagger;\\ \mathbf{W}^l = &\ \mathtt{prox}_{\frac{\rho_2}{\eta_W^l}\|\cdot\|_*}(\mathbf{W} + \frac{\Gamma_W^l}{\eta_W^l}); \mathbf{W}^c = \mathtt{prox}_{\frac{1}{\eta_W^c}c_2}(\mathbf{W} + \frac{\Gamma_W^c}{\eta_W^c});\\ \Gamma_W^c = &\ \Gamma_W^c - \eta_W^c(\mathbf{W} - \mathbf{W}^c); \Gamma_W^l = \Gamma_W^l - \eta_W^l(\mathbf{W} - \mathbf{W}^l).\end{aligned} \tag{14}$$

Phase II: Robust Underlying Tensor Recovery. In this second phase, we update the low-rank tensor $\{\mathbf{X}_k\}$ which is the underlying clean and completed tensor, and the sparse tensor $\{\mathbf{E}_k\}$ which is the corrupted tensor, as well as the Lagrangian dual variable $\{\Gamma_O^k\}$.

Update $\{\mathbf{X}_k\}$*:* It amounts to

$$\begin{aligned}\mathbf{X}_k = \underset{\mathbf{X}_k}{\operatorname{argmin}} &\|\mathbf{X}_k - \mathbf{Q}_k\mathbf{H}\mathbf{S}_k\mathbf{V}^\top\|_F^2 - \langle\Gamma_O^k, \mathbf{O}_k - \mathbf{X}_k - \mathbf{E}_k\rangle\\ &+ \frac{\eta_O^k}{2}\|\mathbf{O}_k - \mathbf{X}_k - \mathbf{E}_k\|_F^2,\end{aligned}$$

which has the solution

$$\mathbf{X}_k = \mathbf{Q}_k\mathbf{H}\mathbf{S}_k\mathbf{V}^\top - \Gamma_O^k + \eta_O^k(\mathbf{O}_k - \mathbf{E}_k). \tag{15}$$

Update $\{\mathbf{E}_k\}$*:* The update of $\mathbf{E}_k$ separates into $P_\Omega(\mathbf{E}_k)$ and $P_{\Omega^\perp}(\mathbf{E}_k)$:

$$\mathcal{P}_\Omega(\mathbf{E}_k) = \mathcal{P}_\Omega(\mathtt{prox}_{\frac{\rho_0}{\eta_O^k}\|\cdot\|_1}(\mathbf{O}_k - \mathbf{X}_k - \frac{1}{\eta_O^k}\Gamma_O^k)), \tag{16}$$

where $\mathtt{prox}_{\frac{\rho_0}{\eta_O^k}\|\cdot\|_1}(\cdot)$ is the proximal operator for the ℓ_1-norm, a.k.a. soft-thresholding.

$$\mathcal{P}_{\Omega^\perp}(\mathbf{E}_k) = \mathcal{P}_{\Omega^\perp}(\mathbf{O}_k - \mathbf{X}_k). \tag{17}$$

Update $\{\Gamma_O^k\}$*:* This is Lagriangian dual variable update:

$$\Gamma_O^k = \Gamma_O^k - \eta_O^k(\mathbf{O}_k - \mathbf{X}_k - \mathbf{E}_k). \tag{18}$$

3.4 REPAIR: Algorithm and Complexity

The complete REPAIR algorithm is summarized in Algorithm 1. The following theorem summarizes the computational complexity of Algorithm 1.

Algorithm 1 The complete REPAIR algorithm

Input: Input tensor $\mathcal{O}$; Model parameters ρ_0-ρ_3; Optimization parameters η's; Interpretability constraint types c_1, c_2, c_3; Initial rank estimation R.

1: **while** Not reach convergence criteria **do**
2: %% Phase I begins
3: **while** Not reach inner loop max **do**
4: Update $\{\mathbf{Q}_k\}$ by eq.(7);
5: Update $\mathbf{H}, \mathbf{V}, \mathbf{W}$-related variables sequantially;
6: **end while**
7: %% Phase II begins
8: Update $\{\mathbf{X}_k\}$ by eq.(15);
9: Update $\{\mathbf{E}_k\}$ by eq.(16)&(17);
10: Update $\{\Gamma_O^k\}$ by eq.(18).
11: **end while**

Output: Phenotype factor matrices $\{\mathbf{U}_k\} = \{\mathbf{Q}_k\mathbf{H}\}, \{\mathbf{S}_k\}, \mathbf{V}$; Recovered tensor $\{\mathbf{X}_k\}$.

Theorem 1. *(Per-iteration computational complexity of **REPAIR** algorithm) For an input tensor $\mathbf{O}_k : \mathbb{R}^{I_k \times J}$, for $k = 1, ..., K$ and initial target rank estimation R, Algorithm 1's per-iteration complexity is $O(3R^2JK)$.*

Proof. REPAIR's per-iteration complexity breaks down as follows: Line 4 costs $O(\min\{R^2I, RI^2\})$, where I denotes the maximum among $\{I_k\}$; Line 5, updating $\mathbf{H}, \mathbf{V}, \mathbf{W}$ costs $O(3R^2JK)$, updating $\mathbf{H}^l, \mathbf{V}^l, \mathbf{W}^l$ costs $O(R^2(R + J + K))$, updating $\mathbf{H}^c, \mathbf{V}^c, \mathbf{W}^c$ costs $O(R(R + J + K))$; Line 8-10 cost $O(4\sum_{k=1}^{K} I_k J)$. As a result, the per-iteration complexity is $O(3R^2JK)$. □

Remark 1. *If the tensor $\{\mathbf{Y}_k\}$ is sparse and sparse MTTKRP [32] is adopted for updating $\mathbf{H}, \mathbf{V}, \mathbf{W}$, the $O(3R^2JK)$-term further reduces to $O\left(3R^2 nnz(\mathbf{Y})K\right)$, where $nnz(\mathbf{Y})$ denotes the maximum number of none-zero columns among $\{\mathbf{Y}_k\}$. In practice, because of large number of patients K, $R \ll J, K$ and $R + J + K \ll nnz(\mathbf{Y})K$, the overall per-iteration complexity is $O(3R^2 nnz(\mathbf{Y})K)$, which is the same as SPARTan and COPA.*

4 EXPERIMENTS

4.1 Experiment Setup

4.1.1 Datasets.

We evaluate REPAIR on two real-world publicly-available temporal EHR datasets: CMS[2] and MIMIC-III[3].

CMS: Centers for Medicare and Medicaid Services (CMS) contains synthesized data of Medicare beneficiaries in 2008 and their claims from 2008 to 2010. We construct a three-mode tensor with patients (along mode-3), diagnosis or ICD9 codes (along mode-2), and clinical visits (along mode-1). Each tensor value $\mathbf{O}_{ijk}$ indicates the number of times a patient k has a diagnosis j during visit i. We keep records of patients with at least 2 hospital visits. The resulting number of patients is 50,000 with 284 features (diagnosis categories) and the maximum number of observations for a patient is 1500. The number of non-zero elements is 49 million. 89% of the non-zero elements are 1, and 11% are 2.

MIMIC-III: The intensive care unit (ICU) dataset is collected between 2001 and 2012. Similar to CMS, we construct the three-mode tensor and keep records of patients with at least 2 hospital visits. We select 202 ICD-9 codes that have the highest frequency as in [23]. The resulting number of patients is 2323 with 202 features (diagnosis codes) and the maximum number of observations for a patient is 41. The number of non-zero entries is 3 million. 96% of non-zero elements are 1, and 4% are 2.

4.1.2 Methods for Comparison.

Since there are no existing robust methods for irregular tensor factorization with missing and erroneous data, we compare with two groups of methods: 1) state-of-the-art irregular tensor factorization methods, which however have no mechanisms to handle missing and erroneous data; 2) we adapt existing robust methods for regular tensor factorization to irregular tensors for comparison.

1) Irregular tensor factorization methods.

- **SPARTan [32]- scalable PARAFAC2**: A recently-proposed methodology for fitting PARAFAC2 on large and sparse data. It does not explicitly address missing or erroneous data.
- **COPA [2]- scalable PARAFAC2 with additional regularizations**: A state-of-the-art irregular tensor factorization method. It further introduces various constraints/regularizations to improve the interpretability of the factor matrices for more meaningful pheonotype extraction.

2) Adapted robust regular tensor factorization methods.

- **CP-WOPT [1] - robust method for regular tensors**: CP-WOPT is a robust method for *regular* tensors which uses a weighted optimization method for CP tensor completion and factorization with incomplete data. To make it work with irregular tensors, we first zero-pad the irregular tensors to aligned ones and then apply CP-WOPT.
- **lrmcR [28] + COPA - robust method for matrix completion**: lrmcR [28] is a robust low-rank matrix completion method. To make it work for irregular tensors, we apply lrmcR to recover the frontal slices one by one and then apply COPA for phenotype extraction.

4.1.3 Implementation details.

REPAIR[4] is implemented in Matlab R2019a and includes functionalities from the Tensor Toolbox[5]. We utilize the Parallel Computing Toolbox of Matlab. For CMS dataset, 30 workers are used; and for MIMIC-III, 4 workers are used. We report the hyper-parameters of REPAIR in the experiment in Table 3. The code of COPA and SPARTan are publicly available at: https://github.com/aafshar/COPA; https://github.com/kperros/SPARTan. For the COPA related methods, we use the same regularizations c_1, c_2, c_3 with REPAIR, as given in Defintion 3.

We evaluate recovery accuracy and robustness of the tensor factorization against various conditions of missing and erroneous values. We empirical study the convergence behaviour of all compared methods. In case studies, we evaluate the quality of the factorization matrices (i.e. extracted phenotypes) for downstream analysis via:

[2]https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.html
[3]https://mimic.physionet.org/
[4]https://github.com/Emory-AIMS/Repair
[5]https://www.tensortoolbox.org/

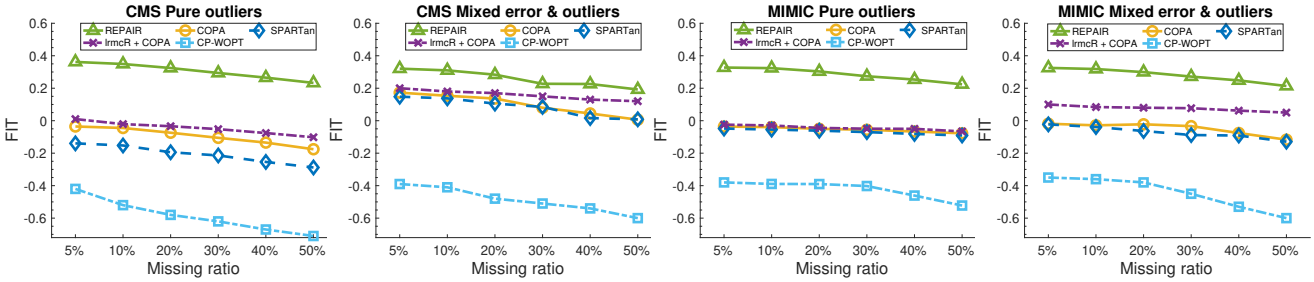


Figure 2: Robustness against varying ratio of missing entries

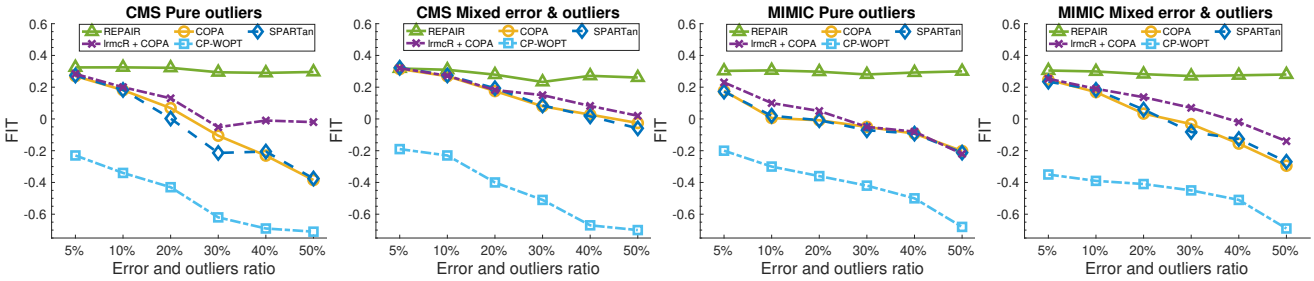


Figure 3: Robustness against varying ratio of erroneous entries

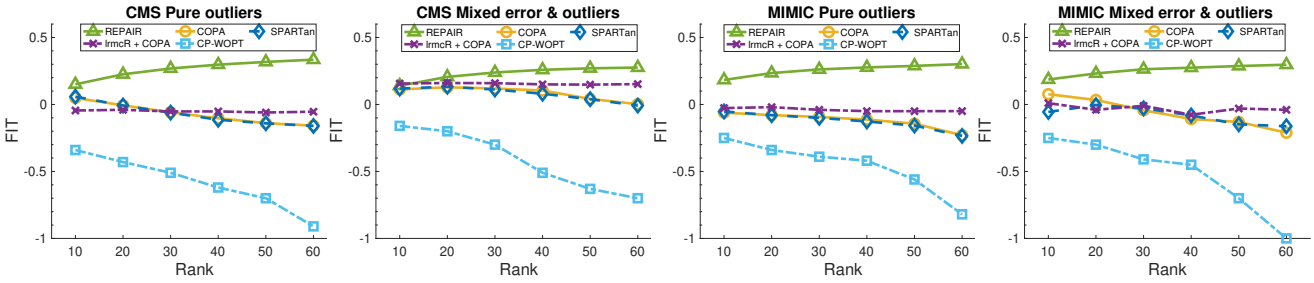


Figure 4: Impact of varying rank estimation

Table 3: Parameters for CMS and MIMIC-III

Parameter	CMS	MIMIC-III
ρ_0	1e-3	1e-3
ρ_1	1e-3	1e-3
ρ_2	1e-4	1e-4
ρ_3	1e-4	1e-4
c_1	253	270
c_3	0.0000085	0.0000085

1) identification of higher-risk patient sub-groups; 2) in-hospital mortality prediction. Finally, we illustrate phenotypes extracted by REPAIR.

4.2 Tensor Factorization Robustness

In order to test the robustness of REPAIR model against missing and error entries, we randomly add missing values and error entries into the two datasets. We design two types of errors. The first is referred as pure outliers, where we randomly pick tensor entries and set their values to be 4, which largely deviates from normal values (1 and 2 in these datasets). The second is mixed error, where we randomly pick certain entries and set their values to be 3 or 4 (outliers) with half probability, and 1 or 2 (normal values but flipped from the original value) with half probability. The original uncorrupted tensor denoted as $\{\mathbf{G}_k\}$ serves as the ground truth. We adopt the $FIT \in (-\infty, 1]$ score [5] as the quality measure (the higher the better):

$$FIT = 1 - \frac{\sum_{k=1}^{K} \|\mathbf{G}_k - \mathbf{U}_k\mathbf{S}_k\mathbf{V}^T\|^2}{\sum_{k=1}^{K} \|\mathbf{G}_k\|^2}. \tag{19}$$

In the following experiment, we run each setting for 5 different random initialization and report the average FIT. When the compared methods' FIT drop below 0 (i.e. fail to recover), we report the averaged highest FIT before the algorithm diverges.

Robustness against Varying Ratio of Missing Entries. We first evaluate the impact of varying missing ratios on the robustness of

Table 4: Basis number for CMS and MIMIC-III

Rank R	CMS	MIMIC-III
10	102	140
20	190	200
30	215	220
40	253	270
50	270	320
60	320	360

the methods with fixed 30% error ratios as Figure 2 shows. we use $R = 40$ and the detailed parameters are shown in Table 3 and 4 in the Appendix. If no error and missing entries are added into data sets, REPAIR, COPA, SPARTAN and lrmcR + COPA methods can achieve similar *FIT* scores around 0.42 (please note that it is a typical FIT range for this task, e.g., [2]). However, the four baselines' *FIT* scores quickly drop as the missing ratio increases, in many cases below 0, which indicates baselines fail to recover the tensor even with small missing ratios. Repair outperforms all methods significantly. lrmcR+COPA performances slightly better than COPA thanks to its completion of the slices. lrmcR + COPA and COPA perform better than SPARTan thanks to its additional temporal constraints. CP-WOPT performs the worst, since it does not address the irregularity of the tensors, even when it explicitly deals with missing data, which indicates the importance of addressing the irregularity. We also observe that pure outlier's performances are often better than mixed error cases, as pure outliers is easier for REPAIR model to separate the error entries.

Robustness against Varying Ratio of Erroneous Entries. We set the missing ratio to be 30%, and change the error ratio from 5% to 50%. Figure 3 shows the *FIT* scores of different methods with respect to varying error ratios for the two data sets under two error cases. With increasing error ratios, four baselines' recovery performance drop dramatically, while REPAIR enjoys a robust performance with an average *FIT* around 0.32.

Impact of Varying Initial Target Rank Estimation. We set missing and error ratios both to 30% and vary the initial rank estimation R. The detailed c_1 (basis function number used by M-spline function for promoting temporal smoothness) for different data sets and various ranks are shown in Table 4. With a higher rank R, the *FIT* of REPAIR slightly increases while always outperforming all other methods as Figure 4 shows. This is because the low-rank regularization function is able to iteratively decrease the target rank during the optimization (e.g. by soft-thresholding the singular values) and make it approach the optimal one.

4.3 Convergence Comparison.

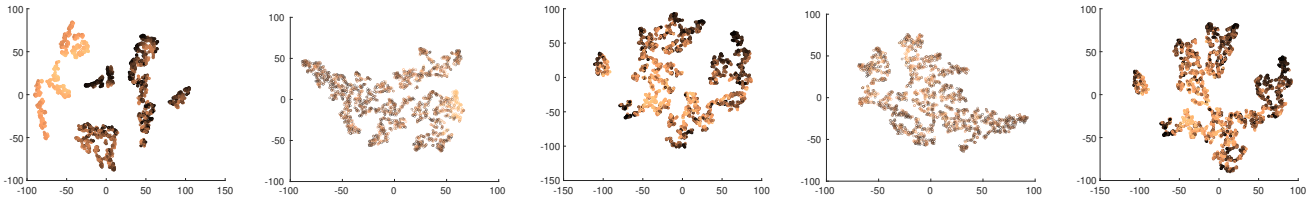
Figure 6 shows the convergence comparison of REPAIR, SPARTan, COPA, lrmcR + COPA, CP-WOPT on CMS with missing ratio 10% and mixed error ratio 20% (under this setting all algorithms can recover the tensor without failure). By Figure 6, REPAIR flats around 9-10 iterations (with a higher FIT score than baselines), while it takes baselines 14-15 iterations. This shows that REPAIR not only enjoys more robust recovery, but also faster convergence.

Table 5: MIMIC-III Phenotypes discovered by REPAIR. The red color corresponds to diagnosis and blue color corresponds to procedures.

Heart failure
Congestive heart failure, unspecified
Atrial fibrillation
Coronary atherosclerosis of native coronary artery
Coronary arteriography using two catheters
Transfusion of packed cells
Left heart cardiac catheterization
Hypertension and hyperlipidemia
Unspecified essential hypertension
Diabetes mellitus without mention of complication
Coronary atherosclerosis of native coronary artery
Other and unspecified hyperlipidemia
Esophageal reflux
Pure hypercholesterolemia
Extracorporeal circulation auxiliary to open heart surgery
Coronary arteriography using two catheters
Single internal mammary-coronary artery bypass
Left heart cardiac catheterization
Kidney disease
Acute kidney failure, unspecified
Hypertensive chronic kidney disease, unspecified
Unspecified essential hypertension
Urinary tract infection, site not specified
Hemodialysis
Venous catheterization for renal dialysis
Transfusion of packed cells
Respiratory failure and sepsis
Acute respiratory failure
Severe sepsis
Atrial fibrillation
Septic shock
Urinary tract infection, site not specified
Insertion of endotracheal tube
Enteral infusion of concentrated nutritional substances
Continuous invasive mechanical ventilation
Closed [endoscopic] biopsy of bronchus
Arterial catheterization
Percutaneous abdominal drainage
Transfusion of packed cells

4.4 Quality of the Extracted Phenotypes: Two Case Studies

The previous experiments show the robustness of REPAIR in terms of how well the factorization matrices (i.e. the extracted phenotypes) recover the ground truth tensor under the *FIT* metric. In this subsection, our goal is to evaluate how meaningful and useful the extracted phenotypes are. Table 5 illustrates the first set of phenotypes extracted by REPAIR when $R = 4$ given corrupted MIMIC-III data, which shows the correlations between diagnosis and procedures related to coronary disease. We next show quantitatively how the phenotypes can be used for various downstream analysis. We



(a) REPAIR (b) SPARTan (c) COPA (d) CP-WOPT (e) lrmcR + COPA

Figure 5: tSNE visualization of patient representations learned by REPAIR, SPARTan, COPA, CP-WOPT, and lrmcR+COPA. Each point represents a patient, the color corresponding to the weight of the "oncological conditions" phenotype (lighter means higher weight).

Method	REPAIR	SPARTan	COPA	CP-WOPT	lrmcR + COPA
Higher-risk Cluster Average Mortality Rate	68.79%	59.86%	60.03%	59.60%	60.55%
Lower-risk Cluster Average Mortality Rate	49.91%	59.13%	58.92%	59.43%	58.45%
Difference	**18.88%**	0.83%	1.11%	0.5%	2.1%

Table 6: Summary of Average Mortality Risk of the higher-risk cluster, lower-risk cluster, and their difference. The two clusters are obtained by k-means clustering ($k = 2$). REPAIR can achieve 18.88% difference, which has the best discriminative capability among all compared methods, under the setting of adding 30% erroneous and 30% missing entries.

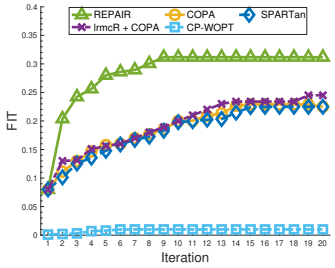


Figure 6: Convergence comparison of REPAIR, SPARTan, COPA, lrmcR + COPA, CP-WOPT

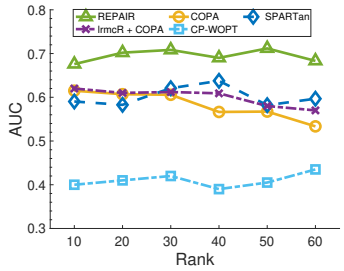


Figure 7: In-hospital mortality prediction in AUC. REPAIR outperforms 17% in terms of prediction performance comparing to the best baseline method lrmcR + COPA

use MIMIC-III for this set of experiments and set both missing and error ratios to 30%.

Identification of Higher-risk Patient Subgroups. The low-dimensional patient representations of PARAFAC2 are effective in distinguishing between higher and lower mortality risk patients [31]. We attempt to test if REPAIR can identify higher-risk patient subgroups if the data contains erroneous and missing entries. The k-th row of patient-by-phenotypes matrix $\mathbf{W} \in \mathbb{R}^{k \times R}$ contains the diagonal of $\mathbf{S}_k$, which indicates importance membership of patient k in each of the phenotypes. We select the largest-variance column among $\mathbf{S}_k$, which is called the "oncological conditions" phenotype. We set $R = 4$, and use the tSNE [36] software to reduce 4-dimensional vectors to 2-dimensional space, and color each point corresponding to the weight of the "oncological conditions" phenotype (lighter means higher weight). As Figure 5 shows, REPAIR can successfully split the patients into two sub-groups while the baselines fail to distinguish the patients.

We perform clustering using K-means (with $k = 2$) on the tSNE result. For the clusters learned by REPAIR, higher risk cluster (corresponding to the left light sub group in Figure 5a) and the lower-risk cluster (corresponding to the right dark sub group in Figure 5a) are 68.79%, 49.91% respectively. We summarize the average mortality risk of the higher-risk cluster, lower-risk cluster, and their difference in Table 6. REPAIR can achieve 18.88% difference, which has the best discriminative capability among all compared methods. In addition, our 18.88%-difference is comparable to the 21%-difference reported in [31], which is the journal extension of the SPARTan algorithm [32], and has a clinical expert's endorsement. Because of the extra error and missing entries, our setting is more challenging than [31]. In sum, it shows our method is robust enough to achieve clinical meaningful result comparable to [31].

In-hospital Mortality Prediction. We also measure REPAIR's phenotype extraction quality under missing and error entries by the predictive power of the discovered phenotypes. A logistic regression

model is trained using the patients' membership indicator S_k as features, which is then utilized for predicting in-hospital mortality. We use five 70-30 train-test splits and evaluate the model using the area under the receiver operating characteristic curve (AUC). As Figure 7 shows, the average score of lrmcR + COPA is 0.605, which performs best among four baselines. REPAIR's average score is 0.703, and offers a 17% prediction performance improvement when compared to lrmcR + COPA, which verifies the robustness and usefulness of the extracted phenotypes.

5 CONCLUSION

We have proposed the REPAIR method for robust irregular tensor factorization and completion with potential missing and erroneous values. It is built on two major contributions: an effective low-rank regularization function specific to PARAFAC2 structure and a two-phase joint optimization for iterative factorization and clean tensor recovery. Extensive experiments have demonstrated that REPAIR can robustly extract meaningful phenotypes from missing and erroneous inputs. In the future, we plan to investigate different loss functions to further enhance the recovery performance and also different types of missing data (in addition to Missing Completely at Random (MCAR) in this paper).

6 ACKNOWLEDGMENT

We sincerely thank all anonymous reviewers for their constructive comments. This work is supported by National Science Foundation (NSF) BigData award IIS-1838200, National Science Foundation (NSF) RAPID award CNS-2027783, Georgia Clinical Translational Science Alliance under National Institutes of Health (NIH) CTSA award UL1TR002378, and National Institute of Health (NIH) award 1K01LM012924-01.

REFERENCES

[1] Evrim Acar, Daniel M Dunlavy, Tamara G Kolda, and Morten Mørup. 2011. Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems* (2011).
[2] Ardavan Afshar, Ioakeim Perros, Evangelos E Papalexakis, Elizabeth Searles, Joyce Ho, and Jimeng Sun. 2018. COPA: Constrained PARAFAC2 for Sparse & Large Datasets. In *CIKM*.
[3] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* (2011).
[4] Rasmus Bro. 1997. PARAFAC tutorial and applications. Chemom Intell Lab Syst. *Chemometrics and Intelligent Laboratory Systems* (1997).
[5] Rasmus Bro, Claus A. Andersson, and Henk A. L. Kiers. 1999. PARAFAC2âĂŤPart II. Modeling chromatographic data with retention time shifts. *Journal of Chemometrics* (1999).
[6] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* (2010).
[7] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. 2011. Robust principal component analysis? *JACM* (2011).
[8] J Douglas Carroll and Jih-Jie Chang. 1970. Analysis of individual differences in multidimensional scaling via an N-way generalization of âĂIJEckart-YoungâĂİ decomposition. *Psychometrika* (1970).
[9] Andrzej Cichocki, Rafal Zdunek, Anh-Huy Phan, and Shun-ichi Amari. 2009. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation.*
[10] Silvia Gandy, Benjamin Recht, and Isao Yamada. 2011. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems* (2011).
[11] Donald Goldfarb and Zhiwei Qin. 2014. Robust low-rank tensor recovery: Models and algorithms. *SIAM J. Matrix Anal. Appl.* (2014).
[12] Richard A Harshman. 1972. PARAFAC2: Mathematical and technical notes. *UCLA working papers in phonetics* (1972).
[13] Richard A Harshman et al. 1970. Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multimodal factor analysis. (1970).
[14] Nathaniel Helwig. 2013. The Special Sign Indeterminacy of the Direct-Fitting Parafac2 Model: Some Implications, Cautions, and Recommendations for Simultaneous Component Analysis. *Psychometrika* (2013).
[15] Frank L. Hitchcock. 1927. The Expression of a Tensor or a Polyadic as a Sum of Products. *Journal of Mathematics and Physics* (1927).
[16] Joyce Ho, Joydeep Ghosh, and J. Sun. 2014. Extracting Phenotypes from Patient Claim Records Using Nonnegative Tensor Factorization. https://doi.org/10.1007/978-3-319-09891-3_14
[17] Joyce C Ho, Joydeep Ghosh, Steve R Steinhubl, Walter F Stewart, Joshua C Denny, Bradley A Malin, and Jimeng Sun. 2014. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics* (2014).
[18] George Hripcsak and David J. Albers. 2013. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association : JAMIA* (2013).
[19] Kejun Huang, Nicholas D Sidiropoulos, and Athanasios P Liavas. 2016. A flexible and efficient algorithmic framework for constrained matrix and tensor factorization. *IEEE Transactions on Signal Processing* (2016).
[20] Peter Jensen, Lars Jensen, and SÃ¥ren Brunak. 2012. Mining electronic health records: Towards better research applications and clinical care. *Nature reviews. Genetics* (2012).
[21] Misha E Kilmer, Karen Braman, Ning Hao, and Randy C Hoover. 2013. Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. *SIAM J. Matrix Anal. Appl.* (2013).
[22] Misha E Kilmer and Carla D Martin. 2011. Factorization strategies for third-order tensors. *Linear Algebra Appl.* (2011).
[23] Yejin Kim, Robert El-Kareh, Jimeng Sun, Hwanjo Yu, and Xiaoqian Jiang. 2017. Discriminative and Distinct Phenotyping by Constrained Tensor Factorization. *Scientific Reports* (2017).
[24] Tamara Kolda and Brett Bader. 2009. Tensor Decompositions and Applications. *SIAM Rev.* (2009).
[25] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. 2012. Tensor completion for estimating missing values in visual data. *IEEE Transactions on pattern analysis and machine intelligence* (2012).
[26] Yuanyuan Liu, Fanhua Shang, Licheng Jiao, James Cheng, and Hong Cheng. 2015. Trace Norm Regularized CANDECOMP/PARAFAC Decomposition with Missing Data. *IEEE Transactions on Cybernetics* (2015).
[27] Canyi Lu, Jiashi Feng, Wei Liu, Zhouchen Lin, Shuicheng Yan, et al. 2019. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).
[28] Canyi Lu, Jiashi Feng, Shuicheng Yan, and Zhouchen Lin. 2017. A unified alternating direction method of multipliers by majorization minimization. *IEEE transactions on pattern analysis and machine intelligence* 40, 3 (2017), 527–541.
[29] Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. 2014. Square deal: Lower bounds and improved relaxations for tensor recovery. In *ICML*.
[30] Neal Parikh, Stephen Boyd, et al. 2014. Proximal algorithms. *Foundations and Trends® in Optimization* (2014).
[31] Ioakeim Perros, Evangelos E. Papalexakis, Richard Vuduc, Elizabeth Searles, and Jimeng Sun. 2019. Temporal phenotyping of medically complex children via PARAFAC2 tensor factorization. *Journal of Biomedical Informatics* (2019).
[32] Ioakeim Perros, Evangelos E Papalexakis, Fei Wang, Richard Vuduc, Elizabeth Searles, Michael Thompson, and Jimeng Sun. 2017. SPARTan: Scalable PARAFAC2 for large & sparse data. In *KDD*.
[33] James O Ramsay et al. 1988. Monotone regression splines in action. *Statistical science* 3, 4 (1988), 425–441.
[34] Bernardino Romera-Paredes and Massimiliano Pontil. 2013. A New Convex Relaxation for Tensor Completion. In *NIPS*.
[35] Ledyard R Tucker. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika* (1966).
[36] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* (2008).
[37] Yichen Wang, Robert Chen, Joydeep Ghosh, Joshua C Denny, Abel Kho, You Chen, Bradley A Malin, and Jimeng Sun. 2015. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *KDD*.
[38] Kun Xie, Can Peng, Xin Wang, Gaogang Xie, and Jigang Wen. 2017. Accurate recovery of internet traffic data under dynamic measurements. In *IEEE INFOCOM*.