Robust Irregular Tensor Factorization and Completion for Temporal Health Data Analysis

Yifei Ren
yifei.ren2@emory.edu
Emory University

Li Xiong
lxiong@emory.edu
Emory University

Jian Lou∗
jian.lou@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 Robust tEmporal PARAFAC2 method for IRregular 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.

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:

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
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 Robust Empirical PARAFAC2 method for Robust tensor factorization and completion (c.f. Figure 1), which is the first robust irregular tensor recovery method. Given each patient input data $O_k$ with erroneous and missing entries, REPAIR performs RLTM to separate out the erroneous entries $E_k$ from the underlying clean and completed components $X_k$, and uses the clean tensor to form a common low rank space for PARAFAC2 based candidate phenotype extraction, i.e. $X_k \approx U_kS_kV^\top$. 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 underlying tensor by decomposing the original tensor ($O_k$) into the underlying low-rank tensor ($X_k$) and the sparse error tensor ($E_k$) (Fig. 1 blue box); and 2) extract phenotypes by factorizing the clean tensor $X_k = U_kS_kV^\top$ (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.

## 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 $O = \{O_k\} \in \mathbb{R}^{k \times J}$ (c.f. leftmost tensor in Figure 1) with 3 modes, where each frontal slice $O_k$ 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 $\{O_k\}$ 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 $O$, RLTM separates it into an underlying clean and completed tensor $X$ and an error tensor $E$. In practice, the clean part is often low-rank while the erroneous part is sparse. Thus, RLTM imposes a low-rank regularization function $\|\cdot\|_r$ and a sparsity regularization function $\|\cdot\|_0$ on $X$ and $E$, correspondingly:

$$\argmin_{X,E} \|X\|_r + \rho_0\|E\|_0, \quad s.t. \ P_{\Omega}(O) = P_{\Omega}(X + E),$$  \tag{1}$$

where $\Omega$ is the index set of non-missing entries and $P_{\Omega}$ keeps entries in $\Omega$ and zeros out others (i.e., missing entries), $\rho_0$ 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
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 \( (\text{rank}(\mathcal{X}_{(1)}), \text{rank}(\mathcal{X}_{(2)}), \text{rank}(\mathcal{X}_{(3)})) \). CP decomposes \( \mathcal{X} \in \mathbb{R}^{I \times J \times K} \) into the sum of \( R \) rank-
one tensors by \( \mathcal{X} = \sum_{r=1}^{R} A(\cdot, r) \circ B(\cdot, r) \circ C(\cdot, r) \), where \( A, B, C \) are
factorization matrices and the smallest \( R \) to achieve such decompo-
sion is defined to be the rank \( R^e \) of \( \mathcal{X} \) under CP model. It is difficult
to accurately estimating \( R^e \) 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 func-
tions 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 \( \{\mathcal{X}_k\} \)
is formalized below [24]:

**Definition 1.** (Classic PARAFAC2 model)

\[
\begin{align*}
\min_{\{U_k\}, \{S_k\}, V} \sum_{k=1}^{K} \frac{1}{2} \|X_k - U_k S_k V^T\|_F^2, \\
\text{s.t. } U_k = Q_k H_k Q_k^T, S_k = I_k, \text{ diag}(S_k) \text{ is diagonal, where } Q_k \in \mathbb{R}^{I_k \times R} \text{ is orthogonal, } I_k \in \mathbb{R}^{R \times K} \text{ is the identity matrix and } R \text{ is the target rank of the \text{PARAFAC2} decomposition.}
\end{align*}
\]

For temporal EHR data, the factorization matrices have the fol-
lowing interpretation:

\(\triangleright U_k \in \mathbb{R}^{I_k \times R}\) contains the temporal evolution for patient
\(k\); the \(r\)-th column of \(U_k\) indicates the evolution of the \(r\)-th
phenotype for all \(I_k\) clinical visits for patient \(k\).

\(\triangleright V \in \mathbb{R}^{J \times R}\) reflects the phenotypes. Each non-zero entry
of \(V\) indicates the membership of the corresponding \(j\)-th
medical feature in the \(r\)-th phenotype.

\(\triangleright S_k \in \mathbb{R}^{R \times R}\) is a diagonal matrix with the importance
membership of patient \(k\) in each one of the \(R\) phenotypes. It is
often organized into \(W \in \mathbb{R}^{J \times R}\) with each row of \(W\) com-
pose by the diagonal of \(S_k\), i.e. \(W(\cdot, k) = \text{diag}(S_k)\).

SPARTan [32] scales PARAFAC2 to large temporal EHR pheno-
typing 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 phenotype extraction. For
example, COPA introduces the M-spline constraint [33] to \(U_k\) to
capture the temporal smoothness, non-negative constraint to \(S_k\) to
get positive weight, and sparsity (e.g., \(\ell_1\) norm regularization) to \(V\)
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 \(\{\mathcal{X}_k\}\)’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 \(\{\mathcal{X}_k\}\).
through adding nuclear norm constraints on the internal factorization matrices \( H, V, W \), which are shared by all frontal slices thus capable of capturing cross-slice information.

**Definition 2.** For irregular tensor \( X = \{ X_k \} \approx \text{PARAFAC2}(\{ Q_k \}, H, V, W) \), the low-rank regularization function is defined as

\[
\| X \|_F := \| H \|_2 + \| V \|_2 + \| W \|_2.
\]

(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 \( X_k \) 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 \( X = \{ X_k \} \) and the erroneous tensor \( E = \{ E_k \} \) given the missing and corrupted observation tensor \( O = \{ O_k \} \). Meanwhile, REPAIR decomposes \( \{ X_k \} \) into PARAFAC2 structure. The tensor recovery of \( X \) is enforced by the linear constraint between \( O, X, E \) in eq (4), low-rank regularization for \( X = \{ X_k \} \) and sparsity constraint for \( E \) in the first row of eq (3). The tensor factorization of \( X \) is enforced by the PARAFAC2 loss for \( X \), 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 \( c_1(H) \), non-negativity for \( \{ c_2(S_k) \} \), sparsity for \( c_3(V) \).

**Definition 3. (REPAIR objective function)**

\[
\begin{aligned}
\text{argmin}_{Q_k, H, S_k, V_k} & \sum_{k=1}^{K} \left( \| X_k - U_k S_k V_k^T \|^2_F + \rho_0 \| P_{\Omega} (E_k) \|_1 \right) + \\
& \text{(low-rankness for } X) \quad \text{smoothness} \quad \text{nonnegativity} \quad \text{sparsity} \\
& \rho_1 \| H \|_2 + \rho_2 \| V \|_2 + \rho_3 \| W \|_2 + c_1(H) + \sum_{k=1}^{K} c_2(S_k) + c_3(V),
\end{aligned}
\]

(3)

\[
\text{subject to } O_k = U_k S_k V_k^T, \quad X_k = S_k \quad \text{for } k = 1, ..., K.
\]

(4)

**Phase I: Approximated PARAFAC2.** In the first phase, we update the factorization matrices \( \{ Q_k \}, H, V, W \) with \( \{ X_k \} \) and \( \{ E_k \} \) fixed, which can be intuitively seen as decomposing the latest recovered tensor \( X_k \) into PARAFAC2. We then observe that it is enough to run PARAFAC2 by one iteration in this phase to make it easier to deal with.

**Update \( Q_k \):** To update \( Q^k \), we need Lemma 1 below:

**Lemma 1.** The Orthogonal Procrustes problem is:

\[
Q^k = \arg\min_{Q_k} \| QA - B \|^2_F,
\]

\[
Q^k Q^k = I,
\]

which has the closed-form solution: \( Q^k = \text{PZ}^T \), where \( \text{PZ} = \text{svd}(BA^T) \) and \( \text{svd}(\cdot) \) is singular value decomposition.

When applied to the update of \( Q_k \), with other factors fixed, we have

\[
Q_k = \arg\min_{Q_k} \| X_k - Q_k HS_k V_k^T \|^2_F.
\]

(6)

**Update \( H \):** After obtaining \( \{ Q_k \} \), we denote \( Y_k = Q_k^T X_k \) for \( k = 1, ..., K \), and let \( Y \) be the tensor with \( X_k \) being its frontal slice. We then update the matrices \( H, V, W \) alternatively by solving three constrained

### 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 \( \{ Q_k \}, H, V, W \); II) separating the \( X \) and \( E \) from \( O \). For I), we factorize the intermediate (inaccurate) recovered tensor \( X \) by solving an approximated PARAFAC2; for II), we follow standard ADMM to convert the linear constraint of eq.(4) into introducing Lagrangian dual variable \( I^k_\Omega \) 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,

\[
\begin{aligned}
& \sum_{k=1}^{K} \left( \| X_k - Q_k HS_k V_k^T \|^2_F - (I^k_\Omega Q_k - X_k - E_k) \\
& + \frac{\rho_0}{2} \| P_{\Omega} (E_k) \|_1 \right) + \\
& + \frac{\rho_1}{2} \| H \|_2 + \rho_2 \| V \|_2 + \rho_3 \| W \|_2 + (c_1(H) + c_2(W) + c_3(V)),
\end{aligned}
\]

s.t. \( S_k = \text{diag}(W(k,:)) \), \( U_k = Q_k H \), \( Q_k^T Q_k = I \).

(7)
Table 2: Additional Symbols for REPAIR Optimization

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_0, \rho_1, \rho_2, \rho_3$</td>
<td>Balancing hyper-parameters</td>
</tr>
<tr>
<td>$H^f, V^f, W^f$</td>
<td>Auxiliary variable for low-rank constr.</td>
</tr>
<tr>
<td>$\Gamma^f_H, \Gamma^c_H$</td>
<td>Lagrangian dual for low-rank constr.</td>
</tr>
<tr>
<td>$\Gamma^c_V, \Gamma^c_W$</td>
<td>Lagrangian dual for interpretability constr.</td>
</tr>
<tr>
<td>$\eta^f_H, \eta^c_V, \eta^c_W$</td>
<td>Lagrangian constant for interpretability constr.</td>
</tr>
</tbody>
</table>

least squares sub-problems. Due to the symmetry of the three sub-problems, we elaborate the update for $H$ as an example.

$$H = \arg \min_H \|Y_{(1)} - H(V \circ W)\|_F^2 + \rho_1\|H\|_* + c_1(H).$$

We introduce two auxiliary variables $H^f$ and $H^c$ to separate the low-rank and interpretability constraints:

$$\arg \min_{H^f, H^c} \|Y_{(1)} - H(V \circ W)\|_F^2 + \rho_1\|H^f\|_* + c_1(H^c)$$

subject to $H^f = H$, $H^c = H$.

The above can be solved by ADMM after introducing the Lagrangian dual variables $\Gamma^f_H, \Gamma^c_H$ and constants $\eta^f_H, \eta^c_H$, correspondingly:

$$\arg \min_{H^f, H^c} \|Y_{(1)} - H(V \circ W)\|_F^2 + \rho_1\|H^f\|_* + c_1(H^c)$$

subject to

$$H^f = H^f, H^c = H^c$$

To solve it by ADMM, we have the following update sequence for $H^f, H^c, H^f$ and dual $\Gamma^f_H, \Gamma^c_H$:

$$H = \left(\frac{\rho_1}{\eta^f_H} \|H\|_* + \eta^f_H \|H\|_* + c_1(H^f)\right)$$

subject to $H^f = H$, $H^c = H$.

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

$$H^f = \text{prox}_{\frac{\rho_1}{\eta^f_H} \|\cdot\|_*}(H + \frac{\Gamma^f_H}{\eta^f_H}) = \text{PDiag}(\max\{0, \frac{\rho_1}{\eta^f_H} - \frac{\Gamma^f_H}{\eta^f_H}\})Z^T,$$

where $[\text{P, Diag}(\sigma), Z] = \text{svd}(H + \frac{\Gamma^f_H}{\eta^f_H})$.

$$H^c = \arg \min_{H^c} \frac{\eta^c_H}{2} \|H^c - H\|_F^2 - \langle \Gamma^c_H, H^c - H \rangle$$

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

$$H^c = \text{prox}_{\frac{\eta^c_H}{2} \|\cdot\|_*}(H + \frac{\Gamma^c_H}{\eta^c_H}).$$

The Lagrangian dual variables are update as follows:

$$\Gamma^c_H = \Gamma^c_H - \eta^c_W (H - H^c);$$

$$\Gamma^c_V = \Gamma^c_V - \eta^c_V (V - V^c);$$

$$W = \left\{ (Y_{(3)}(H \circ W) + \Gamma^c_W, \eta^c_W) \|W\|_* + \rho_1\|W\|_* + c_1(W) \right\}$$

subject to $W = W^c$.

Phase II: Robust Underlying Tensor Recovery. In this second phase, we update the low-rank tensor $X_k$ which is the underlying clean and completed tensor, and the sparse tensor $E_k$ which is the corrupted tensor, as well as the Lagrangian dual variable $\Gamma^k_O$.

Update $\{X_k\}_k$: It amounts to

$$X_k = \arg \min_{X_k} \|X_k - Q_k H S_k V^T \|_F^2 - \Gamma^k_O, O_k - X_k - E_k$$

$$+ \frac{\eta^k}{2} \|O_k - X_k - E_k\|_F^2,$$

which has the solution

$$X_k = Q_k H S_k V^T - \Gamma^k_O + \frac{\eta^k}{2} (O_k - X_k - E_k).$$

Update $\{E_k\}_k$: The update of $E_k$ separates into $P_{\Omega_k}(E_k)$ and $P_{\Gamma^k_O}$(E_k):

$$P_{\Omega_k}(E_k) = P_{\Omega_k}(\text{prox}_{\frac{\eta^k}{\Gamma^k_O} \|\cdot\|_*}(O_k - X_k - \frac{1}{\eta^k} \Gamma^k_O)),$$

where $\text{prox}_{\frac{\eta^k}{\Gamma^k_O} \|\cdot\|_*}(\cdot)$ is the proximal operator for the $\ell_1$-norm, a.k.a. soft-thresholding.

$$P_{\Gamma^k_O}(E_k) = P_{\Omega_k}(O_k - X_k - E_k).$$

Update $\{\Gamma^k_O\}_k$: This is Lagrangian dual variable update:

$$\Gamma^k_O = \Gamma^k_O - \eta^k (O_k - X_k - E_k).$$
Algorithm 1: The complete REPAIR algorithm

Input: Input tensor $O_k$: Model parameters $\rho_0$, $\rho_3$; Optimization parameters $\eta$; Interplay constraint types $c_1$, $c_2$, $c_3$; Initial rank estimation $R$.

1: \textbf{while} Not reach convergence criteria \textbf{do}
2: \hspace{1em} \% Phase I begins
3: \hspace{1em} \textbf{while} Not reach inner loop max \textbf{do}
4: \hspace{2em} Update $\{Q_k\}$ by eq.(7);
5: \hspace{2em} Update $H, V, W$-related variables sequentially;
6: \textbf{end while}
7: \hspace{1em} \% Phase II begins
8: \hspace{2em} Update $\{X_k\}$ by eq.(15);
9: \hspace{2em} Update $\{\mathbf{e}_k\}$ by eq.(16) & (17);
10: \hspace{2em} Update $\{\mathbf{f}_k\}$ by eq.(18);
11: \textbf{end while}

Output: Phenotype factor matrices $\{U_k\} = \{Q_kH\}, \{S_k\}, V$; Recovered tensor $\{X_k\}$.

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.

**Theorem 1. (Per-iteration computational complexity of REPAIR algorithm)** For an input tensor $O_k : \mathbb{R}^{I_k \times J_k \times F}$, 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 $O(\min(R^2I, R^2J))$, where $I$ denotes the maximum among $\{I_k\}$; Line 5, updating $H, V, W$ costs $O(3R^2JK)$, updating $H^T, V^T, M^T$ costs $O(R^2(R + J + K))$, updating $H^F, V^F, W^F$ costs $O((R + J + K))$; 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 $(Y_k)$ is sparse and sparse MTTKRP [32] is adopted for updating $H, V, W$, the $O(3R^2JK)$-term further reduces to $O(3R^2\text{nnz}(Y)K)$, where $\text{nnz}(Y)$ denotes the maximum number of non-zero columns among $\{Y_k\}$. In practice, because of large number of patients $K$, $R \ll J, K$ and $R + J + K \ll \text{nnz}(Y)K$, the overall per-iteration complexity is $O(3R^2\text{nnz}(Y)K)$, 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\textsuperscript{2} and MIMIC-III\textsuperscript{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 $O_{i,j,k}$ 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 phenotype 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 is implemented in Matlab R2019a and includes functionalities from the Tensor Toolbox\textsuperscript{4}. 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/afshar/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 Definition 3.

---

\textsuperscript{1}https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Files/SynPUFs/DE_Syn_PUF.html

\textsuperscript{2}https://mimic.physionet.org/

\textsuperscript{3}https://github.com/Emory-AIMS/Repair

\textsuperscript{4}https://www.tensorflow.org/
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: 1) identification of higher-risk patient sub-groups; 2) in-hospital mortality prediction. Finally, we illustrate phenotypes extracted by REPAIR.

\[ FIT = 1 - \frac{\sum_{k=1}^{K} \| G_k - U_k S_k V^T \|^2}{\sum_{k=1}^{K} \| G_k \|^2}. \]  

### 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 \( G_k \) serves as the ground truth. We adopt the \( FIT \in (-\infty, 1] \) score [5] as the quality measure (the higher the better):

\[ FIT = 1 - \frac{\sum_{k=1}^{K} \| G_k - U_k S_k V^T \|^2}{\sum_{k=1}^{K} \| G_k \|^2}. \]
WOPT performs the worst, since it does not address the irregularity. With increasing error ratios, four baselines' recovery performance drops dramatically, while REPAIR enjoys a robust performance with an average FIT around 0.32.

**Impact of Varying Initial Target Rank Estimation.** We set the initial rank 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.

**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.

**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

<table>
<thead>
<tr>
<th>Rank R</th>
<th>CMS</th>
<th>MIMIC-III</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>102</td>
<td>140</td>
</tr>
<tr>
<td>20</td>
<td>190</td>
<td>200</td>
</tr>
<tr>
<td>30</td>
<td>215</td>
<td>220</td>
</tr>
<tr>
<td>40</td>
<td>253</td>
<td>270</td>
</tr>
<tr>
<td>50</td>
<td>270</td>
<td>320</td>
</tr>
<tr>
<td>60</td>
<td>320</td>
<td>360</td>
</tr>
</tbody>
</table>

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

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

![Figure 5](image-url)

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

![Figure 6](image-url)

<table>
<thead>
<tr>
<th>Method</th>
<th>REPAIR</th>
<th>SPARTan</th>
<th>COPA</th>
<th>CP-WOPT</th>
<th>lrmcR + COPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher-risk Cluster Average Mortality Rate</td>
<td>68.79%</td>
<td>59.86%</td>
<td>60.03%</td>
<td>59.60%</td>
<td>60.55%</td>
</tr>
<tr>
<td>Lower-risk Cluster Average Mortality Rate</td>
<td>49.91%</td>
<td>59.13%</td>
<td>58.92%</td>
<td>59.43%</td>
<td>58.45%</td>
</tr>
<tr>
<td>Difference</td>
<td>18.88%</td>
<td>0.83%</td>
<td>1.11%</td>
<td>0.5%</td>
<td>2.1%</td>
</tr>
</tbody>
</table>

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.

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 \( W \in \mathbb{R}^{k \times R} \) contains the diagonal of \( S_k \), which indicates importance membership of patient \( k \) in each of the phenotypes. We select the largest-variance column among \( S_k \), 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.
We have proposed the REPAIR method for robust irregular 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).

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 ULTR002378, and National Institute of Health (NIH) award 2T32GM00863. This work is also supported by National Science Foundation at Random (MCAR) in this paper).

REFERENCES


[29] Can Mu, Bo Huang, John Wright, and Donald Goldfarb. 2014. Square deal: Lower bounds and improved relaxations for tensor recovery. In ICML.


