

DOCTORAL THESIS

Learning Phenotypes from Electronic Health Records Using Robust Temporal Tensor Factorization

YIN, Kejing

Date of Award:
2021

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

Abstract

With the widespread adoption of electronic health records (EHR), a large volume of EHR data has been accumulated, providing researchers and clinicians with valuable opportunities to accelerate clinical research and to improve the quality of care by advanced analysis of the EHR data. One approach to transforming the raw EHR to actionable insights is computational phenotyping — the process of discovering meaningful combinations of clinical items, e.g. diagnosis and medications, from the raw EHR data for characterizing health conditions with minimum human supervision. Many data-driven approaches have been proposed to tackle the problem, among which non-negative tensor factorization (NTF) has been shown effective for high-throughput discovery of phenotypes from structural EHR data.

Although great efforts have been made, several open challenges limit the robustness of existing NTF-based computational phenotyping models. (1) The correspondence information between different modalities (e.g., between diagnosis and medication) is often not recorded in EHR data, and existing models rely on unrealistic assumptions to construct input tensors for phenotyping which introduces inevitable errors. (2) EHR data are often recorded over time, presenting serious temporal irregularity: patients have different lengths of stay and the time gap between clinical visits can vary significantly. Existing models are limited in considering the temporal irregularity and temporal dependency, which limits their generalizability and robustness. (3) Heavy missingness is unavoidable in the raw EHR data due to recording mistakes or operational reasons. Existing models mostly do not take the missing data into account and assume that the data are fully observed, which can

greatly compromise their robustness.

In this thesis research study, we propose a series of robust tensor factorization models to address these challenges. First, we propose a hidden interaction tensor factorization (HITF) model to discover the inter-modal correspondence jointly with the learning of latent phenotypes. It is further extended to the multi-modal setting by the collective hidden interaction tensor factorization (cHITF) framework. Second, we propose a collective non-negative tensor factorization (CNTF) model to extract phenotypes from temporally irregular EHR data and separate phenotypes that appear at different stages of the disease progression. Third, we propose a temporally dependent PARAFAC2 factorization (TedPar) model to further capture the temporal dependency between phenotypes by capturing the transitions between them over time. Forth, we propose a logistic PARAFAC2 factorization (LogPar) model to jointly complete the one-class missing data in the binary irregular tensor and learn phenotypes from it. Finally, we propose context-aware time series imputation (CATSI) to capture the overall health condition of patients and use it to guide the imputation of clinical time series.

We empirically validate the proposed models using a number of real-world, large-scale, and de-identified EHR datasets. The empirical evaluation results show that the proposed models are significantly more robust than the existing ones. Evaluated by the clinician, HITF and cHITF discovers more clinically meaningful inter-modal correspondence, CNTF learns phenotypes that better separate early and later stages of disease progression, TedPar captures meaningful phenotype transition patterns, and LogPar also derives clinically meaningful phenotypes. Quantitatively, LogPar and CATSI show significant improvement than baselines in tensor completion and time series imputation, respectively. Besides, HITF, cHITF, CNTF, and LogPar all significantly outperform baseline models in terms of downstream prediction tasks.

Keywords: Computational phenotyping, non-negative tensor factorization, binary tensor completion, PARAFAC2 factorization, time series imputation.

Table of Contents

Declaration	i
Abstract	ii
Acknowledgements	iv
Table of Contents	vi
List of Tables	xi
List of Figures	xiv
Chapter 1 Introduction	1
1.1 Tensor Factorization for Computational Phenotyping	3
1.2 Our Contributions	5
1.3 Thesis Organization	7
Chapter 2 Preliminaries and Related Work	9
2.1 Regular Tensors and CP Factorization	9
2.2 Irregular Tensors and PARAFAC2 Factorization	12
2.3 Review of Related Work	14
Chapter 3 Learning Inter-Modal Correspondence and Phenotypes from Multi-Modal Electronic Health Records	18
3.1 Introduction	18
3.2 Proposed Framework	21

3.2.1	HITF: Hidden Interaction Tensor Factorization	23
3.2.2	Towards Multiple Modalities: collective Hidden Interaction Tensor Factorization Framework	27
3.2.3	Interpretability-Enhancing Regularizations	30
3.2.4	Learning Algorithms	31
3.3	Experiments	33
3.3.1	Datasets	33
3.3.2	Baselines and Hyperparameter Tuning	33
3.3.3	HITF Discovers Correspondence with Significantly Improved Clinical Meaningfulness	34
3.3.4	cHITF Infers Clinically Relevant Phenotypes	42
3.3.5	cHITF Infers Phenotypes with Improved Predictive Power	49
3.3.6	cHITF Models Modalities with Different Distributions	50
3.3.7	cHITF is Efficient and Scalable	51
3.3.8	Sensitivity Analysis of Hyperparameters	53
3.3.9	Summary of Experiments	55
3.4	Summary	56
Chapter 4	Learning Phenotypes and Dynamic Patient Representa- tions via RNN Regularized Collective Non-negative Ten- sor Factorization	57
4.1	Introduction	57
4.2	Proposed Model	59
4.2.1	Collective Non-Negative Tensor Factorization	60
4.2.2	Incorporating Non-temporal Data Modality	62
4.2.3	RNN-based Temporal Regularization	63
4.2.4	Learning Algorithms	64
4.3	Experiments and Results	66
4.3.1	Data Set	66
4.3.2	Phenotypes	67

4.3.3	Sparsity and Similarity	70
4.3.4	Interpretation of the Dynamic Patient Representations	71
4.3.5	Mortality Prediction Task	73
4.4	Summary	74

**Chapter 5 TedPar: Temporally Dependent PARAFAC2 Factorization
for Phenotype-based Disease Progression Modeling 76**

5.1	Introduction	76
5.2	Temporally Dependent PARAFAC2	79
5.2.1	Notations	79
5.2.2	Framework Overview	79
5.2.3	Phenotype-Transition Mechanism	82
5.2.4	Transitional Under-representation Penalty	84
5.2.5	Learning Algorithms	84
5.3	Experiments	86
5.3.1	Experimental Setup	86
5.3.2	Reconstruction and Generalization	88
5.3.3	Clinical Relevance of Phenotypes and Transition Patterns	93
5.4	Summary	95

**Chapter 6 LogPar: Logistic PARAFAC2 Factorization for Temporal
Binary Data with Missing Values 97**

6.1	Introduction	97
6.2	Background on Low-Rank Completion of Binary Matrix	99
6.3	Proposed Method	100
6.3.1	Logistic PARAFAC2 Factorization	100
6.3.2	Non-negative Positive-Unlabeled Loss	102
6.3.3	Regularization	103
6.3.4	Learning Algorithms	105
6.3.5	Theoretical Analysis	106

6.4	Experiments and Results	110
6.4.1	Datasets	110
6.4.2	Baselines	112
6.4.3	Hyperparameter Setting	112
6.4.4	Tensor Completion	113
6.4.5	Downstream Predictive Tasks	118
6.4.6	Phenotype Case Study	120
6.4.7	Ablation Study	122
6.5	Summary	123

Chapter 7 Context-Aware Time Series Imputation for Multi-Analyte

	Clinical Data	124
7.1	Introduction	124
7.2	Background on Time Series Imputation	126
7.3	Dataset	128
7.4	Methodology	130
7.4.1	Notations	130
7.4.2	Pre-processing: Normalization and Completion by Temporal Decay	131
7.4.3	Context-Aware Recurrent Imputation	132
7.4.4	Cross-Feature Imputation	135
7.4.5	Imputation Fusion	136
7.4.6	Loss Function and End-to-end Training	136
7.5	Experiments and Results	137
7.5.1	Individual Missingness	137
7.5.2	Consecutive Missingness	139
7.5.3	Ablation Study: Effect of the Recurrent and Cross-Feature Components	142
7.6	Summary	144

Chapter 8	Conclusions and Future Work	146
8.1	Summary of the Thesis	146
8.2	Limitations	147
8.3	Future Work	148
Bibliography		149
Curriculum Vitae		160

List of Tables

3.1	Symbols and notations used in Chapter 3.	22
3.2	Search spaces and optimal values of hyperparameters.	35
3.3	Examples of Diagnosis-Medication Correspondence Inferred by HITF and Rubik. The number following each item is the corresponding score inferred. Items in red bold text are annotated by the clinician to be clinically meaningful, and the rest are not meaningful.	40
3.4	Examples of Diagnosis-Lab-Test Correspondence Inferred by HITF and Rubik. The number following each item is the corresponding score inferred. Items in red bold text are annotated by the clinician to be clinically meaningful, that in blue italic text are possibly meaningful, and the rest are not meaningful. “[B]” denotes lab tests for blood.	41
3.5	Three examples of clinically relevant phenotypes inferred by cHITF.	45
3.6	Sparsity and diversity of phenotypes inferred from MIMIC-III dataset.	47
3.7	The AUPRC score for predicting in-hospital mortality of MIMIC-III.	48
3.8	The AUPRC score for predicting in-hospital mortality of eICU.	48
3.9	Performance with different combinations of modalities in MIMIC-III.	51
3.10	Examples of diagnosis-fluid correspondence inferred by cHITF.	52

4.1	Three examples of the learned phenotypes. The rows correspond to diagnoses, abnormal laboratory results and medications respectively, where the numbers between parentheses are the weightings. Due to space limitation, only the first three items are listed.	68
4.2	Three examples of the phenotypes derived by the Rubik model. The rows correspond to diagnoses, abnormal laboratory results and medications respectively. Due to space limitation, only the first three items are listed.	69
4.3	Sparsity and similarity of phenotypes derived by the proposed CNTF model and the baseline Rubik model.	71
5.1	Basic statistics of the dataset used in Chapter 5.	86
5.2	Reconstruction accuracy measured by FIT score over the training set and test set with varying numbers of phenotypes. \bar{R} is the total number of phenotypes. The bold number in each column indicates the best score. TedPar consistently outperforms all baselines over both training and test set, and obtains the smallest gap between training and test set.	89
5.3	Two examples of the background phenotypes.	95
6.1	Basic statistics of the datasets used in Chapter 6.	111
6.2	Hyperparameter setting for different datasets.	111
6.3	Three examples of the phenotypes extracted from the Sutter dataset. The weights inside the parentheses after the phenotype index is the logistic regression coefficient for predicting case patients for heart failure. “Dx” denotes for diagnoses and “Rx” denotes for medications.	121
6.4	The completion performance for the ablation study of LogPar in Sutter dataset, measured by PR-AUC. “Uni.” and “Smth.” are abbreviations for uniqueness regularization and temporal smoothness regularization, respectively.	122

7.1	Basic characteristics of the 13 analytes extracted, where RSD is the relative standard deviation of the empirical mean. The empirical mean and the RSD of the training set and the test set are listed separately, and the interquartile range and the missing rates are for the training set.	129
7.2	Symbols and notations used throughout Chapter 7.	130
7.3	Experimental results for individual missingness, the original shared task in the ICHI19 data analytics challenge. The performance scores of the 3D-MICE model were provided by the challenge organizers. CATSI obtained 10.6% relative improvement over 3D-MICE and BRITS.138	
7.4	Missing rates of the newly generated consecutive missing datasets, where m indicates the number of consecutive missing values.	140
7.5	Imputation performance for consecutive missingness, measured using the normalized root mean square deviation (n RMSD).	141
7.6	Experimental results of the imputation performance using CATSI and the two components separately. Experiments are done with the individual missing dataset as in Section 7.5.1. The numbers in bold indicate the best performance for the corresponding analyte, and those in italics indicate the second-best ones. The performance is measured using normalized Root Mean Square Deviation (n RMSD).	143

List of Figures

- 2.1 Non-negative Tensor Factorization for Computational Phenotyping: Each resulting rank-one tensor is interpreted as a phenotype, where its entries with non-zero values are extracted as the definition of the phenotype. For example, phenotype 1 consists of two diagnoses: *Cardiac dysrhythmias* and *heart failure*, and two medications: *Metoprolol* and *Furosemide*. “Dx” denotes diagnosis and “Rx” denotes medications. 12

- 2.2 PARAFAC2 model for computational phenotyping: The input is a collection of binary matrices, with each of them corresponding to a patient. They have the same number of columns representing diseases, but different numbers of rows representing clinical visits. Value 1s in those matrices indicate confirmation of disease while value 0 means either the absence of the disease or missing diagnosis. 13

- 3.1 Real examples of the diagnosis-medication correspondence from MIMIC-III dataset: Each row denotes a diagnosis (Dx1/Dx2/Dx3) and the “1/0” value next to it indicates if the diagnosis is present or not. Each column denotes a medication (Rx1/Rx2/Rx3) and the number underneath each medication denotes the amount of prescribed medications. (a) Adopting equal-correspondence strategy. (b) Correspondence inferred by the proposed HITF model, which is more reasonable. 20

3.2	The building block: Hidden Interaction Tensor Factorization (HITF), illustrated in a third-order example with medications and diagnoses. Only the marginalization along the medication mode and the diagnoses mode are known, leaving the interactions totally unobserved. The hidden interaction tensor are assumed to be drawn from some distribution p parameterized by the CP factorization of the hidden interaction tensor.	25
3.3	Overview of the cHITF framework, illustrated with an concrete example with three hidden interaction tensors and four modalities. Generally the number of interaction tensors and the modalities involved in each tensor can be determined by the nature of the data and the problem flexibly. Each hidden interaction tensor may follow different distributions, and marginalize to a subset of the observed matrices. The factor matrices (<i>i.e.</i> , the phenotypes and the patient representations) are used to reconstruct the hidden interaction tensor, with the factor matrix corresponding to the same modality being forced to be the same. Finally, the learned patient representations can be used as features for subsequent tasks.	28
3.4	The process of quantitatively evaluating the inter-modal correspondence, illustrated with an example of medications for diabetes. Step 1: We gather all the correspondence items from all models and randomly shuffle them. Step 2: We present the items to clinicians for blind scoring. There are three options: 0 for “not clinically relevant”, 1 for “possibly clinically relevant”, and 2 for “clinically relevant”. Step 3: We compute the quality score of each model by taking weighted sum of the score given by the clinician weighted by the weighting produced by each model.	37

3.5	The meaningfulness score of medications and lab tests inferred by HITF and baselines. The inter-modal correspondence inferred by HITF are significantly better than baselines. “Dx” is the abbreviation of diagnosis, and the percentage inside the parentheses denotes the frequency of the corresponding diagnosis. The ten diagnoses listed in the figure are as follows. Dx1 : Cardiac dysrhythmias, Dx2 : Heart failure, Dx3 : Other forms of chronic ischemic heart disease, Dx4 : Diabetes mellitus, Dx5 : Disorders of fluid electrolyte and acid-base balance, Dx6 : Bacterial infection in conditions classified elsewhere and of unspecified site, Dx7 : Iron deficiency anemias, Dx8 : Chronic bronchitis, Dx9 : Arterial embolism and thrombosis, Dx10 : Symptoms involving nervous and musculoskeletal systems.	39
3.6	Quantitative evaluations of the clinical relevance of the phenotypes inferred. cHITF outperforms the baselines significantly by having 18 phenotypes annotated as relevant by the clinical expert.	43
3.7	Running time of cHITF and baselines on the MIMIC-III subset with modalities of diagnosis, medication and lab test.	52
3.8	Sensitivity of the prediction performance, sparsity and similarity to different hyperparameters. (a) : Number of phenotypes; (b) : Weighting parameter of the elastic net regularization (γ); (c) : Weighting parameter of the angular regularization (β); (d) : The ratio of ℓ_1 term in the elastic net regularization (α); (e) : The threshold parameter in the angular regularization (θ_n); (e) : The variance in the Gaussian distribution for input fluid modality (σ^2). The red line represent the in-hospital mortality prediction performance, the dotted blue line represent the sparsity, and the solid blue lines represent the cosine similarity. The vertical dotted black lines represent the final optimal value we used in the experiments.	54

4.1	The framework of the proposed model. A 3 rd order <i>time-labtest-medication</i> tensor is constructed for each patient, and all of the temporal tensors are factorized with the phenotype definitions being shared across all the patients. RNN-based regularization is introduced to model the time dependency of the dynamic patient representations. Another tensor model called HITF [81] is also incorporated to allow non-temporal modalities to be utilized.	59
4.2	Visualization of three examples of the dynamic patient representations. Each row corresponds to a phenotype, and the grey level indicates the weighting of the phenotype at different time points (normalized by the maximum value of each row). The definitions of phenotype 1, 4 and 9 are given in Table 4.1.	72
4.3	Prediction accuracy of in-hospital mortality at different time.	73
5.1	Overview of TedPar: We approximate the irregular tensor of case patients with temporally dependent and independent components. The former corresponds to the portion of the data with progression patterns relevant to the target diseases, characterized by a set of “target” phenotypes and the transition probabilities between them over time. The latter accounts for the remaining portion of the data which are “irrelevant” to the target diseases, characterized by a set of “background” phenotypes which are further forced to approximate also the temporal matrix of control patients to help better separate the clinical features that are relevant and irrelevant to the target disease progression.	80
5.2	The FIT score measured over clean raw dataset with varying noise ratio in the test set.	92
5.3	An example of the phenotype transition bases.	92

5.4	Two examples of the learned phenotype transitions. The items in blue with “[DX]” tags are diagnosis codes and that in red with “[RX]” tags are medications.	94
6.1	Tensor completion performance with different target ranks. PR-AUC is used as the evaluation metric as the tensors are binary. LogPar consistently outperforms all baselines for all datasets, and is more robust to overfitting when the target rank is large.	114
6.2	Tensor completion performance with different missing ratio. PR-AUC is used as the evaluation metric as the tensors are binary. LogPar consistently outperforms all baselines for all level of missingness and is robust to large missing rate.	115
6.3	Performance of heart failure prediction using the Sutter dataset.	119
6.4	Performance of mortality prediction using the MIMIC-III dataset.	119
7.1	The framework of Context-Aware Time Series Imputation (CATSI) model, consisting of two major components: the context-aware recurrent component and the cross-feature component. The final imputation is produced by fusing the recurrent imputations and the cross-feature imputations using a fusion layer.	126
7.2	Architecture of the context-aware recurrent component based on a bidirectional LSTM. The vector \mathbf{r} denotes the global context vector learned from the entire time series input. At each time step, the global context vector, together with the observations before and after, are used to learn the hidden states of the forward and backward RNNs. The recurrent imputation is then obtained by combining the two hidden states learned.	132
7.3	Loss function for the recurrent component during training.	143