# Extracting Phenotypes from Patient Claim Records Using Nonnegative Tensor Factorization

Joyce C. Ho[1], Joydeep Ghosh[1], and Jimeng Sun[2]

[1] Electrical and Computer Engineering Department
The University of Texas at Austin, Austin, TX, 78712 USA
[2] College of Computing
Georgia Institute of Technology, Atlanta, GA 30332

**Abstract.** Electronic health records (EHRs) are becoming an increasingly important source of patient information. Unfortunately, EHR data do not always directly and reliably map to medical concepts that clinical researchers need or use. Some recent studies have focused on EHR-derived phenotyping, which aims at mapping the EHR data to specific medical concepts; however, most of these approaches require labor intensive supervision from experienced clinical professionals.

In this paper, we use Limestone, a nonnegative tensor factorization method to derive phenotype candidates from claims data with virtually no human supervision. Limestone represents the interactions between diagnoses and procedures among patients naturally using tensors (a generalization of matrices). The resulting tensor factors are reported as phenotype candidates that automatically reveal patient clusters on specific diagnoses and procedures. To the best of our knowledge, this is the first study that successfully extracts useful phenotypes by applying sparse nonnegative tensor factorization to a large, public-domain EHR dataset covering a broad range of diseases. Our experiments demonstrate the interpretability and the promise of high-throughput phenotypes generated from tensor factorization.

**Keywords:** EHR phenotyping, tensor factorization, dimensionality reduction.

## 1 Introduction

Electronic health records (EHRs), an important source of detailed patient information, are increasingly becoming prevalent within the U.S healthcare system, with federal incentives for meaningful use of EHRs serving as a major driving force. The complexity of the data stored in EHR systems has grown with the widespread adoption of EHRs. EHRs are composed of a diverse array of data, such as structured information (e.g. diagnosis, medications, lab results), molecular sequences, unstructured clinical progress notes, and social network information. Effective integration and efficient analysis of EHRs help physicians make informed clinical decisions; providers improve patient safety; and researchers discover new knowledge and facilitate investigations [1]. While

D. Ślęzak et al. (Eds.): BIH 2014, LNAI 8609, pp. 142–151, 2014.

data-driven approaches are revolutionizing the field of medical informatics [1–3], several formidable challenges arise from the application of EHR data to clinical research. These include: (i) diverse population, where the data cover patients from various providers who use different and incompatible EHR systems; (ii) heterogeneous and noisy information; (iii) sparsely sampled event sequences with varying time scales; (iv) modeling interactions amongst different data sources (types); and (v) reluctance of medical practitioners to act on any recommendations unless they can understand the findings and reconcile them with existing domain knowledge. The interpretability constraints arise because medical professionals are accustomed to reasoning based on concise and meaningful medical concepts, or phenotypes. Recent work has focused on EHR-based phenotyping, a process to map raw EHR data into meaningful medical concepts, Phenotyping approaches learn medically relevant characteristics of the data [4] and is crucial for supporting genome-wide association studies [5].

State of the art phenotype developments rely primarily on approaches that are heuristic, rule, and iterative based, and are a collaborative team effort between clinicians and IT experts [4, 6]. Examples of large-scale phenotyping efforts are typified by the Electronic Medical Records and Genomics (eMERGE) Network [7], which explores the use of EHRs to obtain phenotypic information at multiple medical institutions, and the Observational Medical Outcomes Partnership (OMOP) [8]. However, phenotypes are often disease-centric and the development of a phenotype for a single disease can take months [9]. Thus, data mining and machine learning tools have been leveraged for high-throughput phenotyping, or efficient and automated phenotype extractions to reduce manual development [4, 10]. Yet, current high-throughput methodologies cannot generate large amounts of candidate phenotypes and achieve good performance without human annotated samples [10]. Therefore, two major limitations of existing phenotyping efforts are (i) the need for human annotation of case and control samples, which take substantial time and effort and (ii) the lack of formalized methodology to derive novel phenotypes.

One possible approach for high-throughput phenotyping of EHR data is to use dimensionality reduction techniques [4]. The "ideal" phenotype (i) represents complex interactions between several sources, (ii) is concise and easily understood by a medical professional, and (iii) maps to domain knowledge. Thus, phenotyping can be viewed as a form of dimensionality reduction, where each phenotype forms a latent space [4]. Matrix factorization, a common dimensionality reduction approach, is insufficient as it cannot concisely capture structured EHR source interactions, such as multiple procedures performed to treat a single disease. A more natural transformation is tensor factorization, which utilizes the multiway structure to produce concise and potentially more interpretable results. We recently proposed *Limestone*, a nonnegative tensor factorization model, to simultaneously generate multiple phenotypes from EHR data with minimal human supervision [11] for the problem of characterizing heart failure. Our pilot study extracted 50 phenotypes from Geisinger Health System's EHRs that were

**Table 1.** List of notations used in this paper

| Symbol | Definition |
|---|---|
| $\boldsymbol{\lambda}, \mathbf{a}$ | vector |
| $\mathbf{A}$ | matrix |
| $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{M}}$ | tensor |
| $\boldsymbol{i}$ | tensor element index $(i_1, i_2, \cdots, i_N)$ |
| $x_{\boldsymbol{i}}$ | tensor element at index $\boldsymbol{i}$ |
| $\circ$ | outer product |

evaluated by an experienced cardiologist. The results were extremely promising, as 82% of the phenotypes generally mapped to a medical concept.

This paper briefly describes the Limestone model and evaluates the model on a publicly available, realistic set of claims data covering a much broader range of diseases. Our experimental results demonstrate the conciseness of the resulting phenotypes. Furthermore, we analyze the phenotypes associated with four common chronic disease conditions.

## 2 Preliminaries and Related Work

This section describes the preliminaries of matrix and tensor decomposition and related tensor factorization work. Table 1 provides a key for the symbols used in the paper. For indexing of matrix $\mathbf{A}$, we denote the $(i,j)^{\text{th}}$ element as $a_{ij}$ and the $j^{\text{th}}$ column as $\mathbf{a}_j$.

**Matrix Decomposition.** Matrix factorization (MF) is a common dimensionality reduction approach, which represents the original data using a lower dimensional latent space. Standard MF approaches find two lower dimensional matrices that when multiplied together approximately produce the original matrix, $\mathbf{X} \approx \mathbf{WH}$. Although many matrix decomposition techniques exist, singular value decomposition and nonnegative matrix factorization (NMF) are two common algorithms used to reduce the feature dimension.

**Tensor Decomposition.** A tensor is a generalization of matrices to higher dimensions. Tensor representations are powerful because they can capture relationships for high-dimensional data. A tensor is rank-one if it can be written as follows:

**Definition 1.** *The outer product of $N$ vectors, $\mathbf{a}^{(1)} \circ \mathbf{a}^{(2)} \circ \cdots \circ \mathbf{a}^{(N)}$, produces a rank-one, $N^{th}$ order tensor $\boldsymbol{\mathcal{X}}$ where each element $x_{\boldsymbol{i}} = x_{i_1,i_2,\cdots,i_N} = a^{(1)}_{i_1} a^{(2)}_{i_2} \cdots a^{(N)}_{i_N}$.*

Tensor factorization (decomposition) is a natural extension of matrix factorization and utilizes information from the multiway structure that is lost when modes are collapsed to use matrix factorization algorithms [12, 13]. The CANDECOMP / PARAFAC (CP) [14, 15] model is a common tensor decomposition

and can be viewed as a higher-order generalization of singular value decomposition [16]. The CP model approximates the original tensor $\boldsymbol{\mathcal{X}}$ as a sum of $R$ rank-one tensors and can be expressed as

$$\begin{aligned}\boldsymbol{\mathcal{X}} &\approx \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^{(1)} \circ \ldots \circ \mathbf{a}_r^{(N)} \\ &= [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \ldots; \mathbf{A}^{(N)}]\!].\end{aligned}$$

Note that $[\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \ldots; \mathbf{A}^{(N)}]\!]$ is shorthand notation to describe the CP decomposition, where $\boldsymbol{\lambda}$ is a vector of the weights $\lambda_r$ and $\mathbf{a}_r^{(n)}$ is the $r$th column of $\mathbf{A}^{(n)}$. The CP tensor decomposition has been used for concept discovery [17], network analysis of fMRI data [18], and community discovery [19]. The details of computing the CP decomposition and other tensor decomposition models can be found in [16].

Some domain applications may desire nonnegative components, a higher-order generalization of NMF. Nonnegative tensor factorization (NTF) requires the elements of the factor matrices and the weights to be nonnegative. A broad survey of practical and useful NMF and NTF algorithms can be found in [20]. Our paper will focus on the nonnegative CP alternating Poisson regression (CP-APR) model to fit sparse count data [21]. For convenience, the CP-APR optimization problem is provided:

$$\begin{aligned}\min f(\boldsymbol{\mathcal{M}}) &\equiv \sum_{\boldsymbol{i}} m_{\boldsymbol{i}} - x_{\boldsymbol{i}} \log m_{\boldsymbol{i}} \\ \text{s.t } \boldsymbol{\mathcal{M}} &= [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; ...; \mathbf{A}^{(N)}]\!] \in \Omega \\ \Omega &= \Omega_\lambda \times \Omega_1 \times \cdots \times \Omega_N \\ \Omega_\lambda &= [0, +\infty)^R \\ \Omega_n &= \{\mathbf{A} \in [0,1]^{I_n \times R} \mid ||\mathbf{a}_r||_1 = 1 \ \forall r\},\end{aligned}$$

where $\boldsymbol{\mathcal{M}}$ is the CP tensor factorization that approximates the observed tensor $\boldsymbol{\mathcal{X}}$, $\Omega$ is the sample space of $\boldsymbol{\mathcal{M}}$, and $I_n$ refers to the size of the $n$th mode. Details of the algorithm and model are presented in the paper by Chi and Kolda [21].

## 3 Limestone Overview

Limestone is a tensor factorization model to achieve high-throughput phenotyping from EHR data. The model is an extension of CP-APR to produce concise phenotype definitions for better interpretability. For this paper, we construct a tensor using the count of the co-occurrences between diagnoses and procedures. Thus, each tensor element denotes the number of times a procedure $p$ is performed to treat diagnosis $d$ for patient $p$. This third-order tensor is then approximated using the CP decomposition $\boldsymbol{\mathcal{M}} = [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}]\!]$, shown in Figure 1. The factor matrix for the $n^{\text{th}}$ mode, $\mathbf{A}^{(n)}$, defines the elements from the

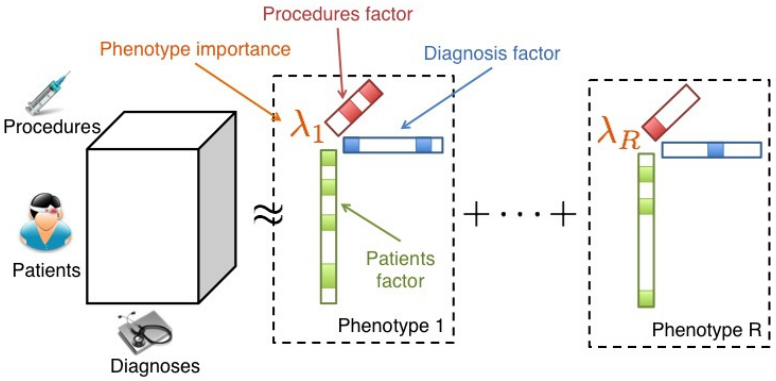



**Fig. 1.** Generating candidate phenotypes from the patient × diagnosis × procedure tensor using CP tensor factorization

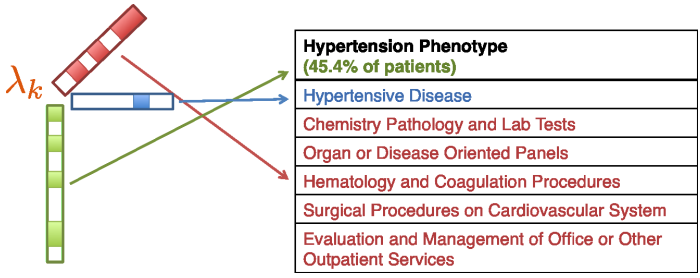



**Fig. 2.** An example of the $k^{\text{th}}$ candidate phenotype produced from the tensor factorization, and the interpretation of the tensor factorization result. The green text, blue, and red text correspond to non-zero elements in the patient, diagnosis, and medication factors, respectively.

mode that comprise the candidate phenotypes. Limestone minimizes the presence of "minuscule and unnecessary" factor components via a hard-thresholding operator [22]. The hard-threshold constraint sets individual factor components $a_{jr}^{(n)}$ that are below a specified threshold ($\gamma_n$) to zero.

We provide an illustrative example of a Limestone phenotype from the claims record data in Figure 2. Given the $k^{\text{th}}$ phenotype, $a_{ik}^{(j)}$ represents the probability of seeing the $i^{\text{th}}$ element in the $j^{\text{th}}$ mode. In our example, hypertensive disease was the only non-zero element in the $k^{\text{th}}$ column of the diagnosis factor matrix while there are 5 non-zero elements in the $k^{\text{th}}$ column of the procedure factor matrix. The percentage of patients with the phenotype is calculated using the percentage of non-zero elements in the $j^{\text{th}}$ column of the patient factor matrix. The candidate phenotype shows that 45.4% of the patients had a non-zero element in the $k^{\text{th}}$ column.

# 4 Experimental Results

## 4.1 Data Description

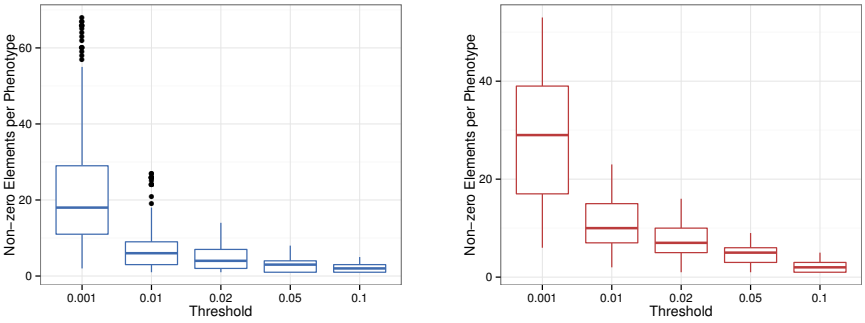
The Centers for Medicare and Medicaid Services (CMS) provides the *CMS Linkable 2008-2010 Medicare Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)*, a publicly available dataset that contains inpatient, oupatient, carrier, and prescription drug event claims in addition to the patient summary files. The claim records have been synthesized from 5% of the 2008 Medicare population, spans 3 years, and is over 100 gigabytes (GB) in size . Although the relationships between some of the variables have been altered to protect the privacy of the beneficiaries, the data can still provide interesting and insightful phenotypes. A detailed description of the data can be found on the CMS website[1]. Our experiments focus on a random subset of 10,000 patients from Sample 1 (CMS released the data in 20 separate samples). The EHR tensor is constructed from the carrier claim records using the diagnosis and procedure codes. Since individual International Classification of Diseases (ICD-9) diagnosis codes and Healthcare Common Procedure Coding System (HCPCS) procedure codes capture fine-grained information, we grouped the codes using the Unified Medical Language System Metathesaurus[2], which contains the source vocabularies for over 150 sources, including ICD-9-CM and HCPCS. Aggregating the individual diagnosis codes and procedure codes results in a constructed tensor that is 10,000 patients by 129 diagnoses by 115 procedures.

## 4.2 Threshold Selection

Limestone uses predefined thresholds for each mode, $\gamma_n$, to zero out "probabilistically unlikely" elements. These thresholds provide a tunable knob to adjust the conciseness of the candidate phenotypes. Domain constraints can be used to determine the threshold values (e.g., a phenotype should only contain a maximum of 3 unique diagnoses). However, we explore the effect of the threshold on the number of non-zero phenotypes along the diagnosis and procedure modes. Figure 3 shows a boxplot of the number of non-zero elements per phenotype based on the various threshold values. Note that a low threshold ($\gamma = 0.001$) results in a large number of elements. As the threshold increases, the phenotypes become more concise and more easily interpretable. Based on the plots, the threshold of 0.05 was chosen to allow for slightly more complex phenotype definitions.

[1] The website URL is
`http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/SynPUFs/DE_Syn_PUF.html`

[2] Information about Metathesaurus is located at
`http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html`




(a) Diagnosis (b) Procedure

**Fig. 3.** The distribution of factor elements along the three CMS tensor modes. Zeros entries are omitted from the plot.

### 4.3 Chronic Disease Phenotypes

The United States spends more than 75% of its medical care cost on the treatment of chronic diseases [23]. Furthermore, 68.4% of the Medicare population suffers from 2 or more chronic diseases [24]. Thus, phenotypes relating to chronic disease factors such as heart failure, diabetes, and arthritis can help medical professionals tailor treatment options based on patient's phenotypes and reduce overall healthcare costs. The dataset provides chronic disease indicators that we will use to identify phenotypes associated with specific chronic diseases[3].

**Table 2.** Two phenotypes related to heart failure. The blue and red colors indicate the diagnosis and procedure elements respectively. Within each type, the elements are ordered in decreasing magnitude.

| **Heart Failure Phenotype 1** (36.7% of patients) |
|---|
| Other forms of heart disease |
| Complications of surgical and medical care |
| Hematology and Coagulation Procs. |
| Eval. and Mgmt. of Office or Other Outpatient Svcs. |
| Surgical Procs. on the Cardiovascular System |
| Chemistry Pathology and Laboratory Tests |
| Cardiovascular Procs. |
| Organ or Disease Oriented Panels |

| **Heart Failure Phenotype 2** (30.9% of patients) |
|---|
| Other forms of heart disease |
| Ischemic heart disease |
| Hospital Inpatient Svcs. |
| Eval. and Mgmt. of Office or Other Outpatient Svcs. |

Table 2 depicts two phenotypes related to heart failure. More than 1 in 3 Medicare patients exhibit the first phenotype while a smaller portion (but still substantial) have medical characteristics typified by the second phenotype. The

[3] A patient's chronic condition flag cannot be perfectly reproduced due to the synthetic claim process used.

**Table 3.** Four phenotypes related to diabetes and arthritis. The blue and red colors indicate the diagnosis and procedure elements respectively. Within each type, the elements are ordered in decreasing magnitude.

| **Diabetes Phenotype 1** (34.8% of patients) |
|---|
| Diseases of other endocrine glands |
| Other metabolic and immunity disorders |
| Eval. and Mgmt. of Office or Other Outpatient Svcs. |
| Surgical Procs. on the Cardiovascular System |
| Ophthalmology Procs. |
| Cardiovascular Procs. |
| Urinalysis Procs. |
| Diagnostic/Screening Processes or Results |

| **Diabetes Phenotype 2** (33.1% of patients) |
|---|
| Diseases of other endocrine glands |
| Chemistry Pathology and Laboratory Tests |
| Organ or Disease Oriented Panels |
| Hematology and Coagulation Procedures |
| Surgical Procs. on the Cardiovascular System |
| Eval. and Mgmt. of Office or Other Outpatient Svcs. |

| **Arthritis Phenotype 1** (29.1% of patients) |
|---|
| Arthropathies and related disorders |
| Physical Medicine and Rehabilitation Procs. |
| Eval. and Mgmt. of Office or Other Outpatient Svcs. |

| **Arthritis Phenotype 2** (38.6% of patients) |
|---|
| Arthropathies and related disorders |
| Rheumatism, excluding the back |
| Eval. and Mgmt. of Office or Other Outpatient Svcs. |
| Surgical Procs. on the Musculoskeletal System |
| Surgical Procs. on the Cardiovascular System |
| Cardiovascular Procs. |
| Hematology and Coagulation Procs. |

second phenotype suggests a higher degree of severity as there is an additional heart disease and it requires hospital inpatient services. The two phenotypes demonstrate the potential ability to derive novel phenotypes via a data-driven approach that could otherwise be difficult and time-consuming.

Table 3 depicts another four chronic-disease phenotypes relating to diabetes and arthritis. There are several other chronic disease phenotypes that were extracted, but due to space constraints are not shown in this paper. Note that these phenotypes shown are concise and easily interpretable. In particular, arthritis phenotype 1 contains just 1 diagnosis and 2 procedures and is exhibited in 29% of the population. The procedures are also consistent with known characteristics of the disease, as arthritis sufferers undergo rehabilitation to strengthen their joints. Similar to Table 2, the four phenotypes also demonstrate the power to automatically capture disease severity. Diabetes phenotype 1 suggests diabetes related complications that require cardiovascular surgery, while the second arthritis phenotype captures patients with multiple chronic conditions.

## 5 Conclusion

This paper shows that Limestone offers a data-driven solution to simultaneously generate multiple phenotypes from a diverse EHR population without expert supervision. The experimental results on 10,000 patient records from the CMS

De-SYNPUF dataset demonstrate the conciseness and interpretability of the tensor derived phenotypes. The phenotypes underscore the promise of Limestone for high-throughput phenotyping with minimal human intervention. Limestone can potentially be used to rapidly characterize, predict, and manage a large number of diseases, thereby promising a novel, data-driven solution that can benefit very large segments of the population. Future work will focus on generalizing the sparse nonnegative tensor factorization to multi-relational tensors [19] to incorporate multiple EHR data sources and examine quasi-Newton methods to improve computational speed of the algorithm [25].

**Acknowledgements.** This research is supported by the Schlumberger Centennial Chair in Engineering; Army Research Office under grant W911NF-11-1-0258; and Department of Defense award under award number 60036907.

## References

1. Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nature Reviews: Genetics 13(6), 395–405 (2012)
2. Greengard, S.: A new model for healthcare. Communications of the ACM 56(2), 17–19 (2013)
3. Savage, N.: Better medicine through machine learning. Communications of the ACM 55(1), 17–19 (2012)
4. Hripcsak, G., Albers, D.J.: Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association 20(1), 117–121 (2012)
5. Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., Basford, M.A., Carrell, D.S., Peissig, P.L., Kho, A.N., Pacheco, J.A., Rasmussen, L.V., Crosslin, D.R., Crane, P.K., Pathak, J., Bielinski, S.J., Pendergrass, S.A., Xu, H., Hindorff, L.A., Li, R., Manolio, T.A., Chute, C.G., Chisholm, R.L., Larson, E.B., Jarvik, G.P., Brilliant, M.H., McCarty, C.A., Kullo, I.J., Haines, J.L., Crawford, D.C., Masys, D.R., Roden, D.M.: Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nature Biotechnology 31(12), 1102–1111 (2013)
6. Newton, K.M., Peissig, P.L., Kho, A.N., Bielinski, S.J., Berg, R.L., Choudhary, V., Basford, M., Chute, C.G., Kullo, I.J., Li, R., Pacheco, J.A., Rasmussen, L.V., Spangler, L., Denny, J.C.: Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. Journal of the American Medical Informatics Association 20(e1), e147–e154 (2013)
7. McCarty, C.A., Chisholm, R.L., Chute, C.G., Kullo, I.J., Jarvik, G.P., Larson, E.B., Li, R., Masys, D.R., Ritchie, M.D., Roden, D.M., Struewing, J.P., Wolf, W.A.: eMERGE Team: The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Medical Genomics 4, 13 (2011)
8. Overhage, J.M., Ryan, P.B., Reich, C.G., Hartzema, A.G., Stang, P.E.: Validation of a common data model for active safety surveillance research. Journal of the American Medical Informatics Association 19(1), 54–60 (2012)

9. Hripcsak, G., Albers, D.J.: Correlating electronic health record concepts with healthcare process events. Journal of the American Medical Informatics Association 20(e2), e311–e318 (2013)
10. Chen, Y., Carroll, R.J., Hinz, E.R.M., Shah, A., Eyler, A.E., Denny, J.C., Xu, H.: Applying active learning to high-throughput phenotyping algorithms for electronic health records data. Journal of the American Medical Informatics Association 20(e2), e253–e259 (2013)
11. Ho, J.C., Ghosh, J., Steinhubl, S., Stewart, W., Denny, J.C., Malin, B.A., Sun, J.: Limestone: High-throughput candidate phenotype generation via tensor factorization. Journal of Biomedical Informatics (accepted)
12. Mørup, M.: Applications of tensor (multiway array) factorizations and decompositions in data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(1), 24–40 (2011)
13. Wang, D., Kong, S.: Feature selection from high-order tensorial data via sparse decomposition. Pattern Recognition Letters 33(13), 1695–1702 (2012)
14. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika 35(3), 283–319 (1970)
15. Harshman, R.A.: Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. UCLA Working Papers in Phonetics 16, 1–84 (1970)
16. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Review 51(3), 455–500 (2009)
17. Kang, U., Papalexakis, E., Harpale, A., Faloutsos, C.: Gigatensor: Scaling tensor analysis up by 100 times-algorithms and discoveries. In: Proceeding of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 316–324. ACM (2012)
18. Davidson, I., Gilpin, S., Carmichael, O., Walker, P.: Network discovery via constrained tensor analysis of fMRI data. In: Proceeding of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). ACM (August 2013)
19. Lin, Y.R., Sun, J., Sundaram, H., Kelliher, A., Castro, P., Konuru, R.: Community discovery via metagraph factorization. ACM Transactions on Knowledge Discovery from Data (TKDD) 5(3) (August 2011)
20. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation. Wiley (2009)
21. Chi, E.C., Kolda, T.G.: On tensors, sparsity, and nonnegative factorizations. SIAM Journal on Matrix Analysis and Applications 33(4), 1272–1299 (2012)
22. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning 4(1) (January 2012)
23. Centers for Disease Control and Prevention (CDC): Chronic diseases at a glance 2009. Technical report, CDC (2009)
24. Lochner, K.A., Cox, C.S.: Prevalence of multiple chronic conditions among Medicare beneficiaries, United State 2010. Preventing Chronic Disease: Public Health Research, Practice, and Policy (2013)
25. Hansen, S., Plantenga, T., Kolda, T.G.: Newton-Based Optimization for Nonnegative Tensor Factorizations. arXiv.org (April 2013)