

# PIVETed-Granite: Computational Phenotypes through Constrained Tensor Factorization

Jette Henderson\*  
University of Texas, Austin

Joyce C. Ho  
Emory University

Bradley A. Malin  
Vanderbilt University

Joydeep Ghosh  
University of Texas, Austin

## Abstract

It has been recently shown that sparse, nonnegative tensor factorization of multi-modal electronic health record data is a promising approach to high-throughput computational phenotyping. However, such approaches typically do not leverage available domain knowledge while extracting the phenotypes; hence, some of the suggested phenotypes may not map well to clinical concepts or may be very similar to other suggested phenotypes. To address these issues, we present a novel, automatic approach called PIVETed-Granite that mines existing biomedical literature (PubMed) to obtain cannot-link constraints that are then used as side-information during a tensor-factorization based computational phenotyping process. The resulting improvements are clearly observed in experiments using a large dataset from VUMC to identify phenotypes for hypertensive patients.

**CCS Concepts** • **Computing methodologies** → **Machine learning**; Factorization methods; Regularization;

**Keywords** machine learning, tensor decomposition, learning with auxiliary information, computational phenotyping

## ACM Reference Format:

Jette Henderson, Bradley A. Malin, Joyce C. Ho, and Joydeep Ghosh. 2018. PIVETed-Granite: Computational Phenotypes through Constrained Tensor Factorization. In *Proceedings of 2018 KDD Workshop on Machine Learning for Medicine and Healthcare (MLMH'18)*. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 1 Introduction

Computational phenotyping is the process of algorithmically deriving cohesive sets of clinical characteristics from collections of patient documentation like electronic health records

\*jette@ices.utexas.edu

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

MLMH'18, August 2018, London, United Kingdom

© 2018 Copyright held by the owner/author(s).

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

(EHRs) [10]. For a set of computational phenotypes to be useful to clinicians, they should be 1) sparse (i.e., have relatively few elements) and 2) diverse (i.e., have few overlapping elements). Most importantly, they should map to clinically relevant concepts. Variations of a tensor factorization method called CANDECOMP/PARAFAC (CP) decomposition have shown potential in deriving computational phenotypes with these characteristics. CP decomposition factors multiway arrays, or tensors, into sets of rank-one components. In the application of computational phenotyping, the non-zero elements of each component can be interpreted as elements of a phenotype. Figure 1 shows examples of computational phenotypes derived through tensor factorization. Ho et al. [7] were the first to show tensor decomposition could be applied to tensors constructed from count data extracted from EHRs to derive phenotypes, a large number of which were clinically relevant. Subsequent models have been developed with the goal of deriving sparse, diverse, and interpretable phenotypes [4, 6, 14]. Wang et al. [14] incorporated guidance information from domain experts into the tensor decomposition process, but obtaining input from domain experts may not always be possible. Henderson et al. [4] introduced a CP model called Granite with angular constraints to encourage diverse phenotypes and sparsity constraints to derive succinct phenotypes but found there was a trade-off between the diversity and the clinical meaningfulness of components.

One possible weakness of using CP decomposition is that there can be noise between and across the modes (i.e., elements appear together that do not belong together). In computational phenotyping, this noise could manifest as a medication and diagnosis co-occurring in a component but not actually having a clinical relationship. This weakness can degrade the interpretability of the fits [4]. Few tensor decomposition methods applied to clinical data have used supervision or domain expertise to increase the number of interpretable components [14]. Like many problems in machine learning, incorporating supervision can be challenging and costly in terms of the time and domain expertise necessary for gathering labels or domain-specific constraints.

In this work, we explore a new proxy for domain-expertise via side information extracted using PIVET [5], a phenotype validation tool. The goal is to increase the number of meaningful components in the CP decomposition process without

<b>Phenotype 10</b> 23.67% total patients (40.26% AMI patients) Hypertensive chronic kidney disease Congestive heart failure (CHF) NOS Nonspecific chest pain diuretics angiotensin converting enzyme inhibitors antihyperlipidemic agents	<b>Phenotype 11</b> 23.67% total patients (33.77% AMI patients) Chronic renal failure [CKD] Essential hypertension diuretics glucose elevating agents antihyperlipidemic agents	<b>Phenotype 16</b> 19.73% total patients (28.57% AMI patients) Atrial fibrillation Essential hypertension calcium channel blocking agents adrenal cortical steroids antiarrhythmic agents diuretics nasal preparations inotropic agents	<b>Phenotype 21</b> 25.28% total patients (28.57% AMI patients) Type 2 diabetes antiarrhythmic agents anxiolytics, sedatives, and hypnotics
---	---	---	---

**Figure 1.** PIVETed-Granitephenotypes derived from a tensor constructed from VUMC patient-level data. These phenotypes have high membership of patients who had at least one myocardial infarction.

human input. PIVET analyzes a corpus of publicly available medical journal articles to create evidence sets for candidate phenotypes. Our method, which we refer to as PIVETed-Granite, involves a novel application of PIVET to provide side information in the form of a cannot-link matrix between the diagnosis and medication modes of a tensor constructed from a set of EHRs. By automatically leveraging available biomedical literature, PIVETed-Granite enables more concise and diverse phenotypes that are discriminative and interpretable.

## 2 PIVETed-Granite Framework

*Background.* We briefly describe tensors and the process of decomposing them but refer the reader to [8] for a more detailed account. A tensor is an  $n$ -way or  $n$ -mode array that is used to represent  $n$ -dimensional relationships. We use CANDECOMP/PARAFAC (CP) decomposition [3] to factor tensors. For a tensor  $\mathcal{X}$  with  $N$  modes, the CP decomposition is

$$\mathcal{X} \approx \sum_{r=1}^R \lambda_r \mathbf{a}_r^{(1)} \circ \dots \circ \mathbf{a}_r^{(N)} = \llbracket \boldsymbol{\lambda}; \mathbf{A}^{(1)}; \dots; \mathbf{A}^{(N)} \rrbracket, \quad (1)$$

where each component of the sum is a rank-one tensor formed by taking the outer product of  $N$  vectors,  $\mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \dots \circ \mathbf{a}_r^{(N)}$ . The representation  $\llbracket \boldsymbol{\lambda}; \mathbf{A}^{(1)}; \dots; \mathbf{A}^{(N)} \rrbracket$  is shorthand notation with the weight vector  $\boldsymbol{\lambda} = [\lambda_1 \dots \lambda_R]$  and the factor matrix  $\mathbf{A}^{(n)} = [\mathbf{a}_1^{(n)} \dots \mathbf{a}_R^{(n)}]$ , where  $\mathbf{a}_r$  denotes the  $r^{\text{th}}$  column of  $\mathbf{A}^{(n)}$ . CP decompositions are usually fit using a loss function that makes assumptions about the underlying distribution that generated the data in the tensor. When CP decomposition is applied to tensors constructed from EHRs, each rank-one tensor can be thought of a phenotype.

*Problem Formulation.* PIVETed-Granite is built on existing nonnegative CP decomposition algorithms that model the observed data using the Poisson distribution [4, 6, 7]. We focus on a 3-mode tensor where the three dimensions are (1) patients, (2) diagnoses, and (3) medications, and each element is a count of the number of times a patient received a diagnosis and medication prescription in a given period of time. An observed tensor,  $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$  is approximated as the sum of  $R$  3-way rank-one tensors  $\mathcal{X} \approx \mathcal{Z} = \llbracket \boldsymbol{\lambda}; \mathbf{A}; \mathbf{B}; \mathbf{C} \rrbracket$ , which are the patient, diagnosis, and medication factor matrices,

respectively. To discourage specified diagnosis and medication pairs from appearing together in the same phenotype, PIVETed-Granite introduces a cannot-link matrix between the diagnosis ( $\mathbf{B}$ ) and the medication ( $\mathbf{C}$ ) factor matrices. The optimization problem for the observed tensor  $\mathcal{X}$  is:

$$f(\mathcal{X}) = \min \left( \sum_i (z_i - x_i \log z_i) \right) \quad (2)$$

$$+ \frac{\beta_1}{2} \sum_{r=1}^R \sum_{p=1}^r \left( \left( \max\{0, \frac{(\mathbf{d}_p)^\top \mathbf{d}_r}{\|\mathbf{d}_p\|_2 \|\mathbf{d}_r\|_2} - \theta_d \} \right)^2 \right) \quad (3)$$

$$+ \frac{\beta_2}{2} \sum_{r=1}^R (\|\mathbf{a}_r\|_2^2 + \|\mathbf{b}_r\|_2^2 + \|\mathbf{c}_r\|_2^2) \quad (4)$$

$$+ \beta_3 \text{trace}(\mathbf{B}^\top \mathbf{M} \mathbf{C}) \quad (5)$$

$$\text{s.t } \mathcal{Z} = \llbracket \boldsymbol{\sigma}; \mathbf{u}_a; \mathbf{u}_b; \mathbf{u}_c \rrbracket + \llbracket \boldsymbol{\lambda}; \mathbf{A}; \mathbf{B}; \mathbf{C} \rrbracket \quad (6)$$

$$\mathbf{d} \in \{\mathbf{a}, \mathbf{b}, \mathbf{c}\} \quad (7)$$

$$\|\mathbf{a}_r\|_1 = \|\mathbf{b}_r\|_1 = \|\mathbf{c}_r\|_1 = 1, \mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r \geq 0 \quad (8)$$

$$\|\mathbf{u}_a\|_1 = \|\mathbf{u}_b\|_1 = \|\mathbf{u}_c\|_1 = 1, \mathbf{u}_a, \mathbf{u}_b, \mathbf{u}_c > 0. \quad (9)$$

For count data, the loss function is KL-divergence (2). An angular penalty term (3) discourages any factors from being too similar, where similarity is defined as the cosine angle between two factor vectors, and an  $l_2$  penalty term controls the growth of the size of the factors (4) (See [4] for details).

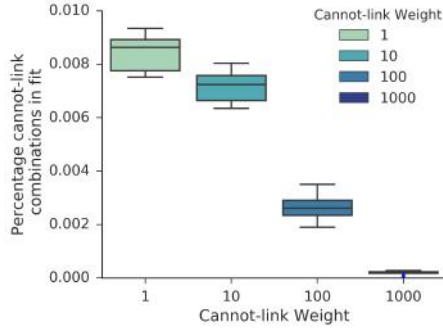
*Incorporating PIVET.* In Equation 5,  $\mathbf{M} \in \mathbf{1}^{I_2 \times I_3}$  is a binary cannot-link matrix defined as follows:

$$\mathbf{M}_{jk} = \begin{cases} 1, & \text{if lift}(b_j, c_k) < \alpha \\ 0, & \text{otherwise} \end{cases}$$

We construct  $\mathbf{M}$  with PIVET. PIVET calculates the lift for each (diagnosis, medication) pair (i.e.,  $b_j, c_k$ ) based on analysis of biomedical journal articles. A lift of much greater than 1 indicates diagnosis  $j$  and medication  $k$  co-occur often and therefore may have a clinical relationship with one another, and a value of 1 or less means diagnosis  $j$  and medication  $k$  do not co-occur often in the corpus and may not have a clinical relationship. In this work, we use  $\alpha = 1$ . The terms in Equation 5 are of the form  $b_{jr} \mathbf{M}_{jk} c_{kr}$ , and only contribute to the objective function if the  $j^{\text{th}}$  diagnosis and the  $k^{\text{th}}$  medication appear in the  $r^{\text{th}}$  component. Since (5) is a soft constraint if there is actually a relationship between  $(b_j, c_k)$  in the data,

they can still appear together in components. However, if the relationship is weak in the data, these elements will be discouraged from appearing together in a phenotype.

We solve for  $\mathbf{Z}$  using gradient descent. We refer the reader to [4] for derivation of the gradients for Equations 2, 3, 4, and [11] for the gradient of Equation 5.



**Figure 2.** Percentage of (diagnosis, medication) cannot-link constraints appearing in the final fit.

### 3 Experiments

*Dataset Description.* To explore the feasibility of using guidance from PIVET, we constructed a tensor from the diagnosis and medication counts of 1622 patients from the Synthetic Derivative (SD), a de-identified EHR database gathered at the VUMC[13]. A panel of domain experts developed sets of characteristics (i.e., billing and medical codes) to identify patients as case and control for a set of diseases[12]. Using these specifications, we included 304 resistant hypertension case patients and 399 resistant hypertension control patients in the tensor. For each case patient, we counted the medication and diagnosis interactions that occurred two years before they received the hypertension diagnosis. For each non-case patient, we counted the interactions that occurred two years before their last interaction with the VUMC. In their raw form, the diagnosis codes (International Classification of Diseases (ICD-9) system) capture a very detailed level of information. We use PheWAS coding to aggregate the diagnosis codes into broader categories[2] and Medical Subject Headings (MeSH) pharmacological terms from the RxClass RESTful API to group the medications into more general categories<sup>1</sup>. These coarser hierarchies produced a tensor with 1622 patients by 1325 diagnoses by 148 medications.

*Quantitative Evaluation.* We compare PIVETed-Granite to two baseline models, Granite and Marble [7]. Table 1 shows fit quality (KL-divergence), sparsity, and diversity measures for the three models ( $R = 30$ ). PIVETed-Granite was the best fit to the data with the lowest KL-divergence, and also resulted in the smallest number of non-zero elements in the diagnosis and medication modes. Additionally, PIVETed-Granite resulted in diagnosis factors that were comparably

diverse to those of Granite and medication factors that were more diverse than Granite. We also evaluate the effect of the cannot-link weight  $\beta_3$  on the percentage of (diagnosis, medication) cannot-link pairs present in the factor matrices. In Figure 2, as  $\beta_3$  increases, the percentage of cannot-link pairs decreases.

Additionally, we evaluated the discriminative capabilities of PIVETed-Granite in a prediction task where the patient factor matrix  $\mathbf{A}$  served as the feature matrix. We compared the performance of PIVETed-Granite, Granite, and Marble using logistic regression to predict which patients were hypertension case and control. The model ran with five 80-20 train-test splits, and the optimal LASSO parameter for the model was learned using 10-fold cross-validation. Table 2 shows the AUC for PIVETed-Granite, Granite, and Marble. The patient factor matrix derived using PIVETed-Granite resulted in the most discriminative model in this task.

*Qualitative Exploration.* To evaluate the effect of the cannot-link matrix  $\mathbf{M}$  on the decomposition process we initialized PIVETed-Granite and Granite fits with the same factors and then examined the differences between the fitted factors. Figure 3 shows one phenotype from each method initialized from the same factors. While the phenotypes are similar to one another, PIVETed-Granite’s characteristics form a more succinct, focused characterization of heart disease complicated with type 2 diabetes. Additionally, the Granite phenotype contains many cannot-link combinations (e.g., (“Fracture of foot”, “antidotes”)) whereas the PIVETed-Granite phenotype does not. The cannot-link constraints seem to result in phenotypes that are descriptive and cohesive.

As a way to qualitatively explore the clinical meaningfulness of the discovered phenotypes we identified patients who experienced acute myocardial infarctions (AMI), which resulted in a cohort of 77 unique patients within the tensor. In Figure 1, we show the phenotypes with the highest proportions of AMI patients. These automatically generated phenotypes seem to give nuanced descriptions of patients who have AMIs. For example, in Phenotype 10 one of the diagnoses is congestive heart failure, which is primarily caused by acute myocardial infarctions [1]. Type-2 diabetes patients (Phenotype 21) are also more likely to experience heart attacks and have more negative outcomes from them [9].

### 4 Discussion and Conclusion

Adding guidance in the form of constraints to computational phenotyping models can help improve the quality of the fit and shows promise in increasing the clinical meaningfulness of derived phenotypes. However, obtaining informative constraints can be difficult and expensive in regard to time and effort required by domain experts. We show how to leverage publicly available information in the form of medical journals to guide the decomposition process to discriminative and interpretable phenotypes. PIVETed-Granite derived phenotypes that were more discriminative, more diverse, and

<sup>1</sup><https://rxnav.nlm.nih.gov/RxClassAPIs.html>

PIVETed-Granite Phenotype (21.64% of patient population)	Granite Phenotype (25.40% of the patient population)
Nonspecific chest pain	Coronary atherosclerosis
Coronary atherosclerosis	Tachycardia NOS
Type 2 diabetes	Rash and other nonspecific skin eruption
Congestive heart failure (CHF) NOS	Type 2 diabetes with renal manifestations
Hyperlipidemia	Chronic pancreatitis
Tachycardia NOS	Pain in joint
angiotensin converting enzyme inhibitors	Vitamin D deficiency
antiarrhythmic agents	Type 2 diabetes
analgesics	Fracture of foot
	Chronic airway obstruction
	Non-Hodgkins lymphoma
	proton pump inhibitors
	angiotensin converting enzyme inhibitors
	antidiabetic agents
	radiologic adjuncts
	vitamins
	antirheumatics
	oral nutritional supplements
	H2 antagonists
	angiotensin receptor blockers
	adrenal cortical steroids
	antidotes

**Figure 3.** Two phenotypes, one derived using PIVETed-Granite (left) and one using Granite (right) where both methods were initialized with the same factor vectors.

**Table 1.** Fit information for phenotypes derived using Marble, Granite, and PIVETed-Granite.

Method	KL-divergence	Average Number of Non-Zeros			Cosine Similarity		
		Patient	Diagnosis	Medication	Patient	Diagnosis	Medication
Marble	2803253.42 (194914.35)	26.72 (1.3)	7.01 (0.37)	8.44 (0.22)	0.07 (0.01)	0.01 (0.01)	0.24 (0.01)
Granite	2311866.35 (27826.92)	70.45 (1.69)	18.21 (1.06)	12.31 (0.31)	0.18 (0.01)	0.02 (0.01)	0.12 (0.01)
PIVETed-Granite	2224824.04 (19758.83)	57.21 (2.65)	5.78 (0.2)	5.89 (0.56)	0.20 (0.02)	0.03 (0.02)	0.05 (0.02)

**Table 2.** AUC for predicting resistant hypertension case patients.

Method	AUC (st. dev.)
Marble	.6656 (.09)
Granite	.7083 (.04)
PIVETed-Granite	<b>.7172 (.01)</b>

sparser than two competing baseline models. In the future, we plan to further analyze the clinical interpretability of the PIVETed-Granite-derived phenotypes and experiment with the threshold used to create the cannot-link constraint matrix. Incorporating cannot-link constraints between modes is a general method that can be applied to many domains. In this application, the quality of the auxiliary information provided by PIVET seems to be high, but in other applications, it may not be. Our next step is to study how to incorporate auxiliary information when that side information is noisy.

## References

- [1] Thomas J Cahill and Rajesh K Kharbanda. 2017. Heart failure after myocardial infarction in the era of primary percutaneous coronary intervention: Mechanisms, incidence and identification of patients at risk. *World journal of cardiology* 9, 5 (2017), 407.
- [2] Joshua C Denny, Lisa Bastarache, Marylyn D Ritchie, and Robert J et al Carroll. 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* 31, 12 (Nov. 2013), 1102–1111.
- [3] Richard A Harshman. 1970. Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multimodal factor analysis. *UCLA Working Papers in Phonetics* 16 (1970), 1–84.
- [4] Jette Henderson, Joyce C. Ho, Abel N Kho, Joshua C Denny, Bradley A Malin, Jimeng Sun, and Joydeep Ghosh. 2017. Granite: Diversified, Sparse Tensor Factorization for Electronic Health Record-Based Phenotyping. In *Fifth IEEE Int. Conf. on Healthc. Inform.* IEEE, 214–223.
- [5] Jette Henderson, Junyuan Ke, C. Joyce Ho, Joydeep Ghosh, and C. Byron Wallace. 2018. Phenotype Instance Verification and Evaluation Tool (PIVET): A Scaled Phenotype Evidence Generation Framework Using Web-Based Medical Literature. *J. Med. Internet Res.* 20, 5 (04 May 2018), e164.
- [6] 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. *J. Biomed. Inf.* 52 (Dec. 2014), 199–211.
- [7] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. 2014. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proc. of ACM Int. Conf. on Knowl. Discov. Data Min.* 115–124.
- [8] Tamara G Kolda and Brett W Bader. 2009. Tensor decompositions and applications. *SIAM Rev.* 51, 3 (2009), 455–500.
- [9] Rodrigo M Lago and Richard W Nesto. 2009. Type 2 diabetes and coronary heart disease: focus on myocardial infarction. *Current diabetes reports* 9, 1 (2009), 73–78.
- [10] Jyotishman Pathak, Abel N Kho, and Joshua C Denny. 2013. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the Am. Med. Inf. Assoc. (JAMIA)* 20, e2 (2013), e206–11.
- [11] Kaare Brandt Petersen and Michael Syskind Pedersen. 2008. The Matrix Cookbook. *Technical University of Denmark* (2008).
- [12] Marylyn D Ritchie, Joshua C Denny, Dana C Crawford, Andrea H Ramirez, Justin B Weiner, et al. 2010. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* 86, 4 (2010), 560–572.
- [13] D. M. Roden, J. M. Pulley, M. A. Basford, G. R. Bernard, E. W. Clayton, J. R. Balser, and D. R. Masys. 2008. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* 84, 3 (2008), 362–369.
- [14] 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 *Proc. of ACM SIGKDD Int. Conf. on Know. Disc. and Data Min.* 1265–1274.