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 (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...
where each component of the sum is a rank-one tensor of \( R \times X \) and medication prescription in a given period of time. An a count of the number of times a patient received a diagnosis on a served data using the Poisson distribution\[4, 6, 7\]. We focus nonnegative CP decomposition algorithms that model the ob-
non each rank-one tensor can be thought of a phenotype.
The distribution that generated the data in the tensor. When CP loss function that makes assumptions about the underlying factor matrix \( A \) and notation with the weight vector \( \lambda \)
formed by taking the outer product of \( N \) vectors, each rank-one tensor.

detailed account. A tensor is an \( n \)-dimensional relationships. We use	
data. For a tensor \( X \) with \( N \) modes, the CP decomposition is

\[
X \approx \sum_{r=1}^{R} \lambda_r a_1^{(r)} \circ \ldots \circ a_N^{(r)} = \llbracket \lambda; A^{(1)}; \ldots; A^{(N)} \rrbracket.
\]

where each component of the sum is a rank-one tensor formed by taking the outer product of \( N \) vectors, \( a_1^{(1)} \circ a_2^{(2)} \circ \ldots \circ a^{(N)} \). The representation \( \llbracket \lambda; A^{(1)}; \ldots; A^{(N)} \rrbracket \) is shorthand notation with the weight vector \( \lambda = [\lambda_1; \ldots; \lambda_R] \) and the factor matrix \( A^{(n)} = [a_1^{(n)}; \ldots; a_R^{(n)}] \), where \( a_r \) denotes the \( r^{th} \) column of \( A^{(n)} \). 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 ob-
erved data using the Poisson distribution\[4, 6, 7\]. We focus on a 3-mode tensor where the three dimensions are (1) pa-
tients, (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, \( X \in \mathbb{R}^{h_1 \times h_2 \times h_3} \) is approximated as the sum of \( R \) 3-way rank-one tensors \( X \approx Z = \llbracket \lambda; A; B; C \rrbracket \), which are the patient, diagnosis, and medication factor matrices, respectively. To discourage specified diagnosis and medica-
tion pairs from appearing together in the same phenotype, PIVETed-Granite introduces a cannot-link matrix between the diagnosis (\( B \)) and the medication (\( C \)) factor matrices. The optimization problem for the observed tensor \( X \) is:

\[
f(X) = \min \left( \sum_{i} (z_i - x_i) \log z_i \right) \]

\[
+ \frac{\beta_1}{2} \sum_{r=1}^{R} \sum_{p=1}^{r} \left( \max(0, \frac{\|d_p\|_2 \|d_r\|_2 - \theta_d)^2 \right) \]

\[
+ \frac{\beta_2}{2} \sum_{r=1}^{R} \left( \|a_r\|_2^2 + \|b_r\|_2^2 + \|c_r\|_2^2 \right) \]

\[
+ \beta_3 \text{trace}(B^TMC) \]

s.t \( Z = \llbracket \sigma; u_a; u_b; u_c \rrbracket + \llbracket \lambda; A; B; C \rrbracket \)

\[
d \in \{a, b, c\} \]

\[
\|a_r\|_1 = \|b_r\|_1 = \|c_r\|_1 = 1, a_r, b_r, c_r \geq 0 \]

\[
\|u_a\|_1 = \|u_b\|_1 = \|u_c\|_1 = 1, u_a, u_b, u_c > 0. \]

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, \( M \in \mathbb{R}^{l_1 \times l_2} \) is a binary cannot-link matrix defined as follows:

\[
M_{jk} = \begin{cases} 
1, & \text{if lift}(b_j, c_k) < \alpha \\
0, & \text{otherwise} 
\end{cases}
\]

We construct \( M \) 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 \( b_j M_{jk} c_k \), and only contribute to the objective function if the \( j^{th} \) diagnosis and the \( k^{th} \) medication appear in the \( r^{th} \) component. Since (5) is a soft constraint if there is actually a relationship between \( b_j, c_k \) in the data,
PIVETed-Granite resulted in diagnosis factors that were comparably more 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 $A$ 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 $M$ 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
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              | 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)        | 18.21 (1.06)      | 5.89 (0.56) | 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)  | 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              | 0.6656 (.09)   |
| Granite             | 0.7083 (.04)   |
| PIVETed-Granite     | 0.7172 (.01)   |

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 “exploratory” 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 Healh. 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 Steinhulh, 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 Know. 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] Jette Henderson, Junyuan Ke, Abel N Kho, Joyce C Denny, Bradley A Malin, and Jimeng Sun. 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. Bailer, and D. R. Mays. 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. 2013. Rubik: Knowledge-guided tensor factorization and completion for health data analytics. In *Proc. of ACM SICKDD Int. Conf. on Know. Disc. and Data Min.* 1265–1274.