# Computational Phenotyping via Scalable Bayesian Tensor Factorization

**Changwei Hu**[1], **Ricardo Henao**[1], **Tahvi Frank**[1], **Shreyas Bhardwaj**[1], **Piyush Rai**[12], **Lawrence Carin**[1]
[1]Duke University, Durham, NC
[2]IIT Kanpur, India
{ch237,r.henao,tahvi.frank,shreyas.bharadwaj,lcarin}@duke.edu,
piyush@cse.iitk.ac.in

## 1 Introduction

Electronic Health Records (EHR) based computational phenotyping utilizes EHR data to extract latent features/factors which represent clinically relevant phenotypes [4, 5, 2, 3, 1]. These latent factors can be used in various downstream analytics, such as identifying at-risk patients, improving prediction of patient morbidity and mortality, or identifying cohorts of patients, medications, diagnoses, etc. (each phenotype may represent a cohort). However, the inherent heterogeneity of the EHR data, usually collected from multiple modalities, such as diagnoses, lab tests, medications, etc., poses significant challenges. Tensor factorization has recently emerged as an attractive way of doing computational phenotyping from such heterogeneous, multimodal EHR data [2, 3]. Here, each dimension of the tensor corresponds to a data modality (e.g., diagnose, lab test, medication, etc.) and each entry within the tensor represents *co-occurrences* (yes/no or counts). For example, a three-way tensor constructed from EHR data representing patients×diagnoses×medications co-occurrences.

In this work, we present a Bayesian latent factor modeling based framework for inferring computational phenotypes from EHR data. In contrast to recent work on tensor factorization for EHR data [2, 3], our framework is not limited to EHR data represented as a single tensor but can seamlessly incorporate a tensor (encoding multiway co-occurrences) as well as additional sources of side-information specified in form of matrices (encoding *pairwise* co-occurrences/relationships between entities within or across modalities), or a vector of outcomes (e.g., denoting patients' hospital admissions or their medical condition). Notably, the computational cost of inference in our model scales in the number of *nonzeros* in the tensor and the associated matrices and outcome vector(s) given as side-information, which is especially appealing because the these objects tend to be highly sparse in the context of EHR data. Moreover, our framework is not limited to count-valued data [2, 3] but can seamlessly handle both count- as well as binary-valued tensor/matrices/vectors. Using a beta-gamma hierarchical construction for the latent factor weights allows us to infer the number of factors (i.e., the number of phenotypes), which is not possible with the existing tensor factorization methods proposed recently for EHR data [2, 3]. Finally, each latent factor along a given tensor dimension represents a distribution (or "topic", as in topic models) over the entities along that dimension, which can be used to rank/cluster the entities within each phenotype, and results in good interpretability.

## 2 Model

We store the co-occurrences in a tensor $\mathcal{Y}$ of size $n_1 \times n_2 \times \cdots \times n_K$, with $n_k$ denoting the size of $\mathcal{Y}$ along the $k^{th}$ mode/dimension of the tensor (where different modes may correspond to patients, diagnoses, medications, lab tests). Due to the space limit, we only describe the case when the only source of information is the tensor; extensions to the case when side-information and/or labels along one or more modes may be available in addition to the tensor are straightforward, and we will discuss these cases briefly towards the end of this section.

We model $\mathcal{Y}$ as a sum of $R$ rank-1 components with a Poisson link function for modeling count data: $\mathcal{Y} \sim \text{Pois}(\sum_{r=1}^{R} \lambda_r \boldsymbol{u}_r^{(1)} \odot \cdots \odot \boldsymbol{u}_r^{(K)})$, where $\boldsymbol{u}_r^{(k)} \sim \text{Dir}(a^{(k)}, \ldots, a^{(k)})$, $\lambda_r \sim \text{Gamma}(g_r, \frac{p_r}{1-p_r})$, $p_r \sim \text{Beta}(c\epsilon, c(1-\epsilon))$, and $\odot$ denotes vector outer product. This construction essentially decomposes $\mathcal{Y}$ into a set of $K$ factor matrices, $\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)}$, where $\mathbf{U}^{(k)} = [\boldsymbol{u}_1^{(k)}, \ldots, \boldsymbol{u}_R^{(k)}]$, for $k = \{1, \ldots, K\}$, denotes the $n_k \times R$ factor matrix associated with mode $k$. The issue to pre-specifying the tensor rank, $R$, is handled via the beta-gamma hierarchical construction for $\lambda_r$, which leads to a shrinkage property [7]. With $R$ sufficiently large, the weight of unnecessary rank-1 components shrinks towards zero, effectively inferring the appropriate tensor rank.

**Side-information and labels, binary observations, and inference:** Our model can be generalized to incorporate side-information or outcomes/labels. For side-information in form of matrices along one or more modes, the latent factors associated with those tensor modes can be shared between the tensor and the matrix, whereas outcomes/labels can be modeled using another layer of regression coefficients associated with the latent factors via a regression latent factor model. Moreover, if the tensor and/or the side-information consists of binary observations, the Poisson link can be replaced by a truncated Poisson link [6]. Our model admits full local conjugacy, using ideas from recent work on Poisson latent factor models [7], which lead to simple inference with excellent computational scalability. The Poisson and truncated Poisson link functions allow our model to scale in the number of nonzero entries in the data [6], which makes it particularly attractive for massive but sparse tensors/matrices/vectors, which are common when working with EHR data. Our inference procedure also naturally extends to streaming data, which allows us to easily incorporate new entities (e.g., new patients) for continuously growing EHR databases. We skip the details here due to the lack of space.

## 3 Experiments

Here we show some preliminary results (not using side-information) with EHR tensor. The data was extracted from a 5-year EHR data (2007-2011) in the care of Durham County residents within Duke University Health System [1]. To narrow our analysis, we focused on a cohort of Type-2 Diabetes Mellitus (T2DM) patients, and identified 16,686 patients in the data. We utilize four modes of data: patients, self-reported medication usage, laboratory tests, and diagnosis/procedure codes. The dataset includes 421 medications (active ingredients), 1,207 types of laboratory tests, 11,825 diagnosis/procedure codes.

Table 1 indicates prominent medications, lab tests and diagnoses/procedures associated with different phenotypes. Due to the limited space, we only list three phenotypes and two entities for medication, lab test and diagnoses/procedure modes. The first phenotype involves gout and chronic kidney disease, which indicates the association between type II diabetes and kidney damage. High blood sugar levels cause damage to kidneys, which can result in hyperuricemia, a buildup of uric acid in the blood that causes gout. The medications associated with this phenotype are allopurinol, to reduce levels of uric acid in the blood, and ketoprofen, to reduce swelling for arthritic patients. The second and third phenotypes correspond to hypertension and heart diseases.

Table 1: Prominent medications, lab tests, and diagnosis/procedure in three phenotypes.

| Med | Test | Diagn/Procedure | Med | Test | Diagn/Procedure | Med | Test | Diagn/Procedure |
|---|---|---|---|---|---|---|---|---|
| Allopurinol | Urea nitrogen | Gout | Amaryl | Glucose | Diabetes mellitus | Amiodarone | Volumn % O2-arterial | Congestive heart failure |
| Ketoprofen | GFR | Chronic kidney dise | Acarbose | Blood | Essential hypertension | Amlodipine | CO2 total-arterial | Diagn ultrasound of heart |

Since each row of a factor matrix is an embeddings/representation of certain entity in the corresponding mode, it can be used to evaluate the distance between entities of different modes. Table 2 associates particular diagnoses/procedures with the top medications by calculating their cosine distances. Amlodipine and amiloride are both drugs that help reduce hypertension, which commonly coexist with diabetes mellitus. Albuterol and zafirlukast are drugs that affect the bronchial smooth muscle to help with lung obstructions. Thus, they are the most commonly associated with asthma. Allopurinol is used to treat excess uric acid in the blood, which is a main cause of gout and kidney disease. Ketoprofen is an anti-inflammatory drug, and one of the conditions it helps treat is gout.

Table 2: Top medications associated with different diagnoses/procedures.

| Diagn/Procedure | Diabetes Mellitus | Asthma | Gout |
|---|---|---|---|
| Med | amlodipine<br>amiloride | albuterol<br>zafirlukast | allopurinol<br>ketoprofen |

Similarly, diagnoses/procedures can also be associated with lab tests using cosine similarities, as shown in Table 3. From the table, we find that glucose test is the most associated with diabetes mellitus, hypertension, and hyperlipidemia. PH arterial blood OR is shown to be connected to cv and arterial catheter, particularly in patients with severe cardiac conditions.

Table 3: Top diagnoses/procedures associated with different lab tests.

| Test | Glucose | PH-arterial blood OR | SPE alpha 2 |
|---|---|---|---|
| Diagn/Proc | diabetes mellitus<br>essential hypertension<br>hyperlipidemia | insert non-tunnel cv cath<br>art cath<br>doppler color flow vel mapping | monoclonal gammopathy<br>bld smear peripheral interp<br>immunoglob typ, ea |

**Acknowledgements** This work is supported by the Duke-Accenture data-analysis project.

## References

[1] R. Henao, Z. Gan, J. Lu, and L. Carin. Deep poisson factor modeling. In *NIPS*, 2015.

[2] J. C. Ho, J. Ghosh, S. Steinhubl, W. Stewart, J. Denny, B. Malin, and J. Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of biomedical informatics*, 52:199–211, 2014.

[3] J. C. Ho, J. Ghosh, and J. Sun. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *KDD*, 2014.

[4] G. Hripcsak and D. Albers. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*, pages 117–121, 2012.

[5] R. Richesson, W. Hammond, M. Nahm, D. Wixted, G. Simon, J. Robinson, A. Bauck, D. Cifelli, M. Smerek, J. Dickerson, R. Laws, R. Madigan, S. Rusincovitch, C. Kluchar, and R. Califf. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the nih health care systems collaboratory. *J Am Med Inform Assoc*, pages 226–231, 2013.

[6] M. Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *AISTATS*, 2015.

[7] M. Zhou, L. A. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and poisson factor analysis. In *AISTATS*, 2012.