Bayesian Double Feature Allocation for Phenotyping with Electronic Health Records

Yang Ni¹,², Peter Müller³, and Yuan Ji⁴,⁵

¹Department of Statistics, Texas A&M University
²Department of Statistics and Data Sciences, The University of Texas at Austin
³Department of Mathematics, The University of Texas at Austin
⁴Department of Public Health Sciences, The University of Chicago
⁵Program of Computational Genomics & Medicine, NorthShore University HealthSystem

September 25, 2018

Abstract

We propose a categorical matrix factorization method to infer latent diseases from electronic health records (EHR) data in an unsupervised manner. A latent disease is defined as an unknown biological aberration that causes a set of common symptoms for a group of patients. The proposed approach is based on a novel double feature allocation model which simultaneously allocates features to the rows and the columns of a categorical matrix. Using a Bayesian approach, available prior information on known diseases greatly improves identifiability and interpretability of latent diseases. This includes known diagnoses for patients and known association of diseases with symptoms. We validate the proposed approach by simulation studies including mis-specified models and comparison with sparse latent factor models. In the application to Chinese EHR data, we find interesting results, some of which agree with related clinical and medical knowledge.

Keywords: Indian buffet process, overlapping clustering, tripartite networks, unsupervised learning, matrix factorization.
1 Introduction

Electronic health records (EHR) data electronically document medical diagnoses and clinical symptoms by the health care providers. The digital nature of EHR automates access to health information and allows physicians and researchers to take advantage of a wealth of data. Since its emergence, EHR has motivated novel data-driven approaches for a wide range of tasks including phenotyping, drug assessment, natural language processing, data integration, clinical decision support, privacy protection and data mining (Ross et al., 2014). In this article, we propose a double feature allocation (DFA) model for latent disease phenotyping. A latent disease is defined as an unknown biological aberration that causes a set of common symptoms for a group of patients. The DFA model is a probability model on a categorical matrix with each entry representing a symptom being recorded for a specific patient. The proposed model is based on an extension of the Indian buffet process (IBP, Ghahramani & Griffiths, 2006). The generalization allows many-to-many patient-disease and symptom-disease relationships and does not require fixing the number of diseases a priori. Existing diagnostic information is incorporated in the model to help identify and interpret latent diseases. DFA can be also viewed as an alternative representation of categorical matrix factorization or as inference for an edge-labeled random network. While recent phenotyping methods (Halpern et al., 2016; Henderson et al., 2017) are mostly performed via supervised learning, the proposed DFA is an unsupervised approach that aims to identify latent diseases.

The proposed approach builds on models for Bayesian inference of hidden structure, including nonparametric mixtures, as reviewed, for example, in Favaro & Teh (2013); Barrios et al. (2013), graphical models (Green & Thomas, 2013; Dobra et al., 2011), matrix factorization (Rukat et al., 2017), and random partitions and feature allocation as discussed and reviewed, for example, in Broderick et al. (2013) or Campbell et al. (2018). We explain below how graphs and matrix factorization relate to the proposed model and inference. The proposed inference is motivated by EHR data that were collected in routine physical examinations for residents in a northeast city in China in 2016. The dataset contains both, information on diagnoses that were recorded by the physicians, as well as symptoms from laboratory test results, such as metabolic and lipid panels etc. The diagnoses are binary variables indicating whether a patient has a disease or not, and includes only some selected
diseases. Data on symptoms are categorical variables with the number of categories depending on the specific type of the symptom. For example, heart rate is divided into three categories, low, normal and high, whereas low density lipoprotein is classified as normal vs abnormal (elevated) levels. The availability of both, diagnostic and symptomatic information provides a good opportunity to detect latent diseases via statistical modeling, that is, to infer additional (latent) disease information in addition to the disease diagnoses that are already included in the data. The proposed DFA model simultaneously allocates patients and symptoms into the same set of latent features that are interpreted as latent diseases.

Many generic methods have been developed for identifying latent patterns, which can be potentially adopted for disease mining. We briefly review related literature. *Graphical models* [Lauritzen, 1996] succinctly describe a set of coherent conditional independence relationships of random variables by representing their distribution as a graph. Conditional independence structure can be directly read off from the graph through the notion of graph separation. Graphical models can reveal certain latent patterns of symptoms. For example, by extracting cliques (maximal fully connected subgraphs) from an estimated graph of symptoms one can identify symptoms that are tightly associated with each other. An underlying disease that is linked to these symptoms may explain the association. However, this approach provides no patient-level inference since the patient-disease relationship is not explicitly modeled and the choice of using cliques rather than other graph summaries remains arbitrary.

*Clustering models* partition entities into mutually exclusive latent groups (clusters). Numerous methods have been developed, including algorithm-based approaches such as k-means, and model-based clustering methods such as finite mixture models [Richardson & Green, 1997; Miller & Harrison, 2017] and infinite mixture models [Lau & Green, 2007; Favaro & Teh, 2013; Barrios et al., 2013] for example).

Partitioning symptoms, similar to graphical models, may discover latent diseases that are related to subsets of symptoms whereas clustering patients suggests latent diseases that are shared among groups of patients. Jointly clustering both symptoms and patients, also known as bi-clustering [Hartigan, 1972; Li, 2005; Xu et al., 2013], allows one to simultaneously learn patient-disease relationships and symptom-disease relationships. The main limitation of clustering methods is the stringent assumption that each patient and symptom is related to exactly one disease. Moreover, most biclustering methods deal with continuous data with
few exceptions (Guo 2013).

**Feature allocation models** (Ghahramani & Griffiths 2006; Broderick et al. 2013), also known as overlapping clustering, relaxes the restriction to mutually exclusive and non-overlapping subsets and allocates each unit to possibly more than one feature. When simultaneously applied to both patients and symptoms, it is referred to as overlapping biclustering or DFA. Like biclustering, most of the existing DFA approaches only handle continuous outcomes. See Pontes et al. (2015) for a recent review and references therein. Only few methods, such as Wood et al. (2006) and Uitert et al. (2008) can be applied to discrete data, but are constrained to binary observations, which is unsuitable for our application with categorical observations. The proposed DFA extends existing methods to general categorical data, automatic selection of the number of features, and incorporation of prior information.

**Latent factor models** assume a low-rank representation of the covariance matrix of a multivariate normal distribution. Latent factor models assume that the variables are continuous and follow independent normal distributions centered on latent factors multiplied by factor loadings. Imposing sparsity constraints (Bhattacharya & Dunson 2011, Ročková & George 2016), latent factor models can be potentially adopted for the discovery of latent causes. However, the assumptions of normality and linear structure are often violated in practice and it is not straightforward to incorporate known diagnostic information into latent factor models. Principal component analysis has the same limitations.

**Matrix factorization** is closely related to factor models but not necessarily constrained to multivariate normal sampling distribution. The most relevant variation of matrix factorization for our application is binary/boolean matrix factorization (Meeds et al. 2007; Zhang et al. 2007; Miettinen et al. 2008) which decomposes a binary matrix into two low-rank latent binary matrices. The proposed DFA can be viewed as categorical matrix factorization which includes binary matrix factorization as a special case.

The contributions of this paper are three-fold: (1) we introduce a novel categorical matrix factorization model based on DFA; (2) we incorporate prior information to identify and interpret latent diseases; and (3) we make inference on the number of latent diseases, patient-disease relationships and symptom-disease relationships.
1.1 Motivating case study: electronic health records

EHR data provides great opportunities for data-driven approaches in early disease detection, screening and prevention. We consider EHR data for \( n = 1000 \) adults aged from 45 to 102 years with median 71 years. The dataset is based on physical examinations of residents in some districts of a city in northeast China, and was collected in 2016. The sample size of \( n=1000 \) corresponds to several weeks’ worth of data. In this paper we focus on developing model-based inference, including full posterior simulation and summaries and therefore focus on a moderate sample size. We will show how meaningful inference about disease discovery is possible with such data. As in any work with EHR data, model-based inference needs to be followed up by expert judgment to confirm the proposed latent diseases and other inference summaries. Once confirmed, inferred disease relationships become known prior information for future weeks. This is how we envision an on-site implementation of the proposed methods.

The data contain blood test results measured on 39 testing items which are listed in Table 1. Figure 1(a) shows the empirical correlation structure of the testing items as a heatmap with green, black, and red colors indicating positive, negligible and negative correlations. With appropriate ordering of the test items, one can see some patterns on the upper-left corner of the heatmap. However, the patterns seem vague and are not easily interpreted. The heatmap of the standardized data is shown in Figure 1(b) with green, black and red colors indicating values above, near and below the average. Next we cluster the data using a K-means algorithm with \( K = 14 \) (the number of latent diseases identified in later model-based inference), applied to both, row and column of the data matrix. The clusters finds some interesting structures. For example, indexing the submatrices in the heatmap by row and column blocks, the values in block (9,9) tend to be above the average whereas the values in the block (1,4) tend to be below the average. However, there are at least two difficulties in interpreting the clusters as latent diseases. Firstly, there is no absolute relationship between the normal range of a testing item and its population average. A deviation from the average does not necessarily indicate an abnormality. Likewise, average values of testing items, especially those related to common diseases such as hypertension, are not necessarily within the normal range. For instance, the mean and the median of systolic blood pressure in our dataset are 147 mm Hg and 145 mm Hg, both of which are beyond the threshold 140 mm Hg.
Hg for hypertension (the high blood pressure values might be related to the elderly patient population). Secondly, the exploratory analysis with K-means does not explicitly model patient-disease relationships and symptom-disease relationships. For example, one may be tempted to interpret each column block as a latent disease. As a consequence, each testing item has to be associated with exactly one disease and the patient-disease relationship is unclear. If instead, we define a latent disease by the row blocks, then each patient has to have exactly one disease and the symptom-disease relationship is ambiguous.

We can slightly improve the interpretability by incorporating prior information. Specifically, each testing item comes with a reference range which we use to define symptoms: a symptom is an item beyond the reference range. In essence, we convert the original data matrix into a ternary matrix which is shown in Figure 1(c). The first difficulty is resolved but the second difficulty remains. For instance, the 6th column seems to suggest a disease with elevated total cholesterol and low density lipoproteins, which is also found in our later analysis with the proposed DFA. However, just as in Figure 1(b), it is hard to judge which blocks meaningfully represent latent disease since patient-disease relationships and symptom-disease relationships are not explicitly modeled. Besides the requirement of specifying the number of clusters, K-means is unsuitable for the task that we are pursuing in this paper. The proposed DFA addresses all these issues, some of which have direct impact in practice. For example, explicit modeling of patient-disease relationships by DFA can be used to estimate the prevalence of latent diseases in the target population, which is helpful for healthcare policymakers. Finally, we emphasize that the gold standard of phenotyping remains the judgment by trained clinicians. The inference from DFA, however, is an important decision tool to facilitate this process, for instance, through the data-assisted personalized diagnosis support system (Section 6.3).

The remainder of this paper is organized as follows. We introduce the proposed DFA model in Section 2 and its alternative interpretations in Section 3. Posterior inference is described in Section 4. Simulation studies and an EHR data analysis are presented in Sections 5 and 6. We conclude this paper by a discussion in Section 7.
Table 1: Blood test items. Acronyms are given within parentheses. Ternary symptoms (after applying the reference range) are in bold face. CV indicates coefficient of variation, “dist” is distribution, “mn” is mean, “ct” is count and “conc” is concentration.

| Item                                    | Description                                      | Acronym            | Item                                    | Description                                      | Acronym            |
|-----------------------------------------|--------------------------------------------------|--------------------|-----------------------------------------|--------------------------------------------------|--------------------|
| alanine aminotransferase (ALT)          |                                                   |                    | aspartate aminotransferase (AST)        |                                                   |                    |
| total cholesterol (TC)                  |                                                   |                    | triglycerides                           |                                                   |                    |
| high density lipoproteins (HDL)         |                                                   |                    | urine pH (UrinePH)                      |                                                   |                    |
| % of monocytes (%MON)                   |                                                   |                    | alpha fetoprotein (AFP)                 |                                                   |                    |
| number of monocytes (#MON)              |                                                   |                    | plateletrit (PCT)                       |                                                   |                    |
| % of eosinophil (%Eosinophil)           |                                                   |                    | basophil ct (#Basophil)                 |                                                   |                    |
| platelet large cell ratio (P-LCR)       |                                                   |                    | platelets                               |                                                   |                    |
| % of granulocyte (%GRA)                 |                                                   |                    | body temperature (BodyTemperature)      |                                                   |                    |
| hemoglobin                              |                                                   |                    | creatinine                              |                                                   |                    |
| glucose                                 |                                                   |                    | diastolic blood pressure (Diastolic)    |                                                   |                    |
| erythrocytes                            |                                                   |                    | hematocrit (HCT)                        |                                                   |                    |
| % of lymphocyte (%LYM)                  |                                                   |                    | mn corpuscular volume (MCV)             |                                                   |                    |
| lymphocyte ct (#LYM)                    |                                                   |                    | granulocyte ct (#GRA)                   |                                                   |                    |
| % of lymphocyte (%LYM)                  |                                                   |                    | % of lymphocyte (%LYM)                  |                                                   |                    |
| % of monocytes (%MON)                   |                                                   |                    | % of monocytes (%MON)                   |                                                   |                    |
| number of monocytes (#MON)              |                                                   |                    | number of monocytes (#MON)              |                                                   |                    |
| % of eosinophil (%Eosinophil)           |                                                   |                    | % of eosinophil (%Eosinophil)           |                                                   |                    |
| platelet large cell ratio (P-LCR)       |                                                   |                    | platelet large cell ratio (P-LCR)       |                                                   |                    |
| % of granulocyte (%GRA)                 |                                                   |                    | % of granulocyte (%GRA)                 |                                                   |                    |
| leukocytes                              |                                                   |                    | creatinine                              |                                                   |                    |
| glucose                                 |                                                   |                    | diastolic blood pressure (Diastolic)    |                                                   |                    |
| hemoglobin                              |                                                   |                    | hematocrit (HCT)                        |                                                   |                    |
| % of lymphocyte (%LYM)                  |                                                   |                    | mn corpuscular volume (MCV)             |                                                   |                    |
| lymphocyte ct (#LYM)                    |                                                   |                    | granulocyte ct (#GRA)                   |                                                   |                    |
| % of lymphocyte (%LYM)                  |                                                   |                    | % of lymphocyte (%LYM)                  |                                                   |                    |
| % of monocytes (%MON)                   |                                                   |                    | % of monocytes (%MON)                   |                                                   |                    |
| number of monocytes (#MON)              |                                                   |                    | number of monocytes (#MON)              |                                                   |                    |
| % of eosinophil (%Eosinophil)           |                                                   |                    | % of eosinophil (%Eosinophil)           |                                                   |                    |
| platelet large cell ratio (P-LCR)       |                                                   |                    | platelet large cell ratio (P-LCR)       |                                                   |                    |
| % of granulocyte (%GRA)                 |                                                   |                    | % of granulocyte (%GRA)                 |                                                   |                    |
| leukocytes                              |                                                   |                    | creatinine                              |                                                   |                    |
| glucose                                 |                                                   |                    | diastolic blood pressure (Diastolic)    |                                                   |                    |
| hemoglobin                              |                                                   |                    | hematocrit (HCT)                        |                                                   |                    |
| % of lymphocyte (%LYM)                  |                                                   |                    | mn corpuscular volume (MCV)             |                                                   |                    |
| lymphocyte ct (#LYM)                    |                                                   |                    | granulocyte ct (#GRA)                   |                                                   |                    |
| % of lymphocyte (%LYM)                  |                                                   |                    | % of lymphocyte (%LYM)                  |                                                   |                    |
| % of monocytes (%MON)                   |                                                   |                    | % of monocytes (%MON)                   |                                                   |                    |
| number of monocytes (#MON)              |                                                   |                    | number of monocytes (#MON)              |                                                   |                    |
| % of eosinophil (%Eosinophil)           |                                                   |                    | % of eosinophil (%Eosinophil)           |                                                   |                    |
| platelet large cell ratio (P-LCR)       |                                                   |                    | platelet large cell ratio (P-LCR)       |                                                   |                    |
| % of granulocyte (%GRA)                 |                                                   |                    | % of granulocyte (%GRA)                 |                                                   |                    |

Figure 1: Heatmap of blood test data. In Panel (a), the correlation structure of the testing items is shown with green, black, and red colors indicating positive, negligible and negative correlations. Diagonal entries are set to 0. In Panels (b) and (c), K-means with $K = 14$ is applied to both row and column of the data matrix. The columns are the testing items and the rows are patients. In Panel (b), the EHR data is standardized. Green, black and red colors represent values above, near and below the average, respectively. In Panel (c), the EHR data is transformed to a categorical matrix based on a trichotomization with respect to a reference range. Green, black and red colors indicate above, within, below the range, respectively.
2 Double Feature Allocation Model

DFA can be applied to any categorical matrix. For simplicity, and in anticipation of the application to EHR, we will describe our model for binary and ternary matrices. Let \( y_{il} \in \{-1, 0, +1\} \) and \( z_{ij} \in \{0, 1\} \) for \( i = 1, \ldots, n \), \( l = 1, \ldots, q \) and \( j = 1, \ldots, p \). In the EHR dataset, \( y_{il} \) denote the observation for patient \( i \) for a symptom that can be naturally trichotomized into low (\( y_{il} = -1 \)), normal (\( y_{il} = 0 \)) and high (\( y_{il} = +1 \)) levels. Similarly, \( z_{ij} \) denotes a symptom that is dichotomized into normal (\( z_{ij} = 0 \)) vs abnormal (\( z_{ij} = 1 \)) levels.

We assume that a patient experiences certain symptoms when he/she has diseases that are related to those symptoms. In other words, the patient-symptom relationships are thought to be generated by two models: a patient-disease model and a symptom-disease model.

2.1 Patient-disease (PD) model

The patient-disease relationships are defined by a binary matrix \( A = (\alpha_{ik}) \in \{0, 1\}^{n \times \tilde{K}} \) with \( \alpha_{ik} = 1 \) meaning patient \( i \) has disease \( k \). We start the model construction assuming a fixed number \( \tilde{K} \) of diseases, to be relaxed later (and we reserve the notation \( \tilde{K} \) for the relaxation).

Conditional on \( \tilde{K} \), \( \alpha_{ik} \) are assumed to be independent Bernoulli random variables, \( \alpha_{ik} \mid \pi_k \sim \text{Ber}(\pi_k) \) with \( \pi_k \) following a conjugate beta prior, \( \pi_k \sim \text{Beta}(m/\tilde{K}, 1) \). Here \( m \) is a fixed hyperparameter. Marginalizing out \( \pi_k \),

\[
p(A) = \prod_{k=1}^{\tilde{K}} \frac{m \Gamma(r_k + m/\tilde{K}) \Gamma(n - r_k + 1)}{\tilde{K} \Gamma(n + 1 + m/\tilde{K})},
\]

where \( r_k = \sum_{i=1}^{n} \alpha_{ik} \) is the sum of the \( i \)-th column of \( A \).

However, the number of latent diseases is unknown in practice and inference on \( \tilde{K} \) is of interest. Let \( H_n = \sum_{i=1}^{n} 1/i \) be the \( n \)-th harmonic number. Next, take the limit \( \tilde{K} \to \infty \) and remove columns of \( A \) with all zeros. Let \( K \) denote the number of non-empty columns. The resulting matrix \( A \) follows an IBP\((m)\) prior (without a specific column ordering) with probability

\[
p(A) = \frac{m^K \exp\{-mH_n\}}{K!} \prod_{k=1}^{K} \frac{\Gamma(r_k) \Gamma(n - r_k + 1)}{\Gamma(n + 1)}.
\]
Since $K$ is taken to be infinity, we do not need to specify the number of latent diseases \textit{a priori}. And with a finite sample size, the number $K$ of non-empty diseases is finite with probability one. The name “Indian buffet process” originates from the culinary metaphor that customers enter an Indian restaurant with an infinite number of dishes. Let $\text{Poi}(\lambda)$ denote a Poisson distribution with rate $\lambda$. The first customer (patient) chooses a $\text{Poi}(m)$ number of dishes (diseases). The $i$th customer takes dish $k$ with probability $r_k/i$ where $r_k$ is the number of earlier customers who have tried dish $k$. Finally, customer $i$ tries a $\text{Poi}(m/i)$ number of new dishes. Let $K$ denote the number of dishes selected by the first $n$ customers. The choices of the first $n$ customers are recorded in a $(n \times K)$ matrix $A$ where $\alpha_{ik} = 1$ if the $i$th customer takes dish $k$. The matrix $A$ is said to follow an IBP. Equation (1) gives the probability distribution of the random matrix after randomly permuting the columns of $A$. The rows of $A$ are exchangeable as the right-hand side of equation (1) does not depend on the row indices of $A$. Due to exchangeability, the conditional probability for $\alpha_{ik} = 1$ is $p(\alpha_{ik} = 1 \mid \alpha_{-i,k}) = r_{-i,k}/n$, provided $r_{-i,k} > 0$ where $\alpha_{-i,k}$ is the $k$th column of $A$ excluding $i$th row, $r_{-i,k}$ is the number of 1’s in $\alpha_{-i,k}$, and the distribution of the number of new features (disease) for each row (patient) is $\text{Poi}(m/n)$.

### 2.2 Symptom-disease (SD) model

A disease may trigger multiple symptoms and a symptom may be related to multiple diseases. The SD model allows both. Let $K$ again denote the number of latent diseases. The SD model inherits $K$ from the PD model. For binary symptoms, we generate a binary $(p \times K)$ matrix $B = [\beta_{jk}]$, $\beta_{jk} \in \{0, 1\}$, from independent Bernoulli distributions, $\beta_{jk} \sim \text{Ber}(\rho)$ with $\rho \sim \text{Beta}(a_\rho, b_\rho)$. Similarly, for ternary symptoms, we generate a $(q \times K)$ ternary matrix $C = [\gamma_{lk}]$, $\gamma_{lk} \in \{-1, 0, 1\}$, from independent categorical distributions $\gamma_{lk} \sim \text{Categ}(\pi)$ with $\pi \sim \text{Dir}(\phi_{-1}, \phi_0, \phi_{+1})$.

Note that $A$, $B$ and $C$ have the same number of columns because they share the same latent diseases. In the language of the Indian restaurant metaphor, each dish (disease) corresponds to a combination of ingredients (symptoms). Some ingredients come at spice levels $\{-1, 1\}$ if selected. We have now augmented model (1) by matching each subset of patients, i.e., each disease, with a subset of symptoms defined in $B$ and $C$. As a result, each disease, or feature, is defined as a pair of random subsets of patients and symptoms,
respectively. We therefore refer to the model as double feature allocation (DFA).

### 2.3 Sampling model

Once we generate the PD and SD relationships, the observed data matrix which records the symptoms for each patient is generated by the following sampling models. Let \( \alpha_i \) denote the \( i \)th row of \( A \), \( \beta_j \) the \( j \)th row of \( B \) as a column vector and \( \gamma_l \) the \( l \)th row of \( C \) as a column vector. We assume conditionally independent Bernoulli distributions for \( z_{ij} \)

\[
    z_{ij} | \alpha_i, \beta_j, W_j, \zeta_j \sim \text{Ber} \left\{ \frac{\exp(\alpha_i W_j \beta_j + \zeta_j)}{1 + \exp(\alpha_i W_j \beta_j + \zeta_j)} \right\},
\]

(2)

with \( W_j = \text{diag}(w_{j1}, \ldots, w_{jK}) \) where \( w_{jk} \) is constrained to be positive so that the probability of experiencing symptom \( j \) for patient \( i \) always increases if a patient has a disease \( k \) that triggers the symptom. The parameter \( \zeta_j \) captures the remaining probability of symptom \( j \) that is unrelated to any disease. One could alternatively include the weight already in \( B \) (and \( C \)). We prefer separating the formation of the random subsets which define the features and the weights which appear in the sampling model.

Similarly, we assume conditionally independent categorical distributions for \( y_{il} \). Let \( \text{Cat}(\pi_1, \ldots, \pi_p) \) denote a categorical distribution with probabilities \( \pi_1, \ldots, \pi_p \) for \( p \) outcomes. Also let \( \gamma_l^+ = I(\gamma_l = +1) \) and \( \gamma_l^- = I(\gamma_l = -1) \) with \( I(\cdot) \) being the element-wise indicator function. We assume

\[
    y_{il} | \alpha_i, \gamma_l, W_l^-, W_l^+, \eta_l^+, \eta_l^- \sim \text{Cat} \left( M e^{\alpha_i W_l^- \gamma_l^- + \eta_l^-}, M, M e^{\alpha_i W_l^+ \gamma_l^+ + \eta_l^+} \right),
\]

(3)

with \( M \) being the normalization constant, \( W_l^- = \text{diag}(w_{l1}^-, \ldots, w_{lK}^-) \) and \( W_l^+ = \text{diag}(w_{l1}^+, \ldots, w_{lK}^+) \), where \( w_{lk}^-, w_{lk}^+ \) are also constrained to be positive, and \( \eta_l^- \) and \( \eta_l^+ \) have interpretations similar to \( \zeta_j \).

We complete the model by assigning weakly informative priors on hyperparameters, \( \zeta_j, \eta_l^-, \eta_l^+ \sim N(0, \tau^2) \) with \( \tau = 100 \) and \( w_{jk}, w_{lk}^-, w_{lk}^+ \sim \text{Gamma}(1, \tau_w) \) with variance \( \tau_w^2 = 100 \).
2.4 Incorporating prior knowledge

Available diagnostic information is easily incorporated in the proposed model. We fix the first $K_0$ columns of $A$ to represent available diagnoses related to $K_0$ known diseases. Similarly, known SD relationships are represented by fixing corresponding columns of $B$ or $C$.

A simple example with 11 patients and 6 ternary symptoms is shown in Figure 2 for illustration. There are 4 diseases, each represented by one color (corresponding to the dashed blocks inside the matrix in Figure 2). Importantly, patients and symptoms can be linked to multiple diseases. For example, patient 9 has both blue and green diseases, and symptom 4 can be triggered by either the red or the green disease. Available prior information is incorporated in this example. For instance, if patients 9, 10, 11 are diagnosed with the blue disease, they will be grouped together deterministically (represented by the blue solid line on the side). Likewise, if the yellow disease is known to lead to symptoms 1 and 6, we fix them in the model (represented by the yellow solid lines on the top).

![Figure 2: Illustration of DFA for ternary symptoms. The lines on the side/top indicate the grouping of patients/symptoms by diseases. Each disease is represented by one color. Dashed lines are latent whereas solid lines are known.](image-url)
3 Alternative Interpretations

The proposed DFA is closely related to matrix factorization and random networks. We briefly discuss two alternative interpretations of DFA for the case of the observed data being ternary. Binary outcomes are a special case of ternary outcomes; generalization to more than three categories is straightforward. We use the same toy example as in Section 2.4 to illustrate the alternative interpretations.

**Categorical matrix factorization (CMF).** DFA can be viewed as a CMF. Merging $W$ and $\gamma$, model (3) probabilistically factorizes an $(n \times q)$ categorical matrix $Y$ into an $(n \times K)$ low-rank binary matrix $A$ and an $K \times q$ low-rank nonnegative matrix $D = (\delta_{kl})$ where $K < \min(n, q)$, $\delta_{kl} = w_{lk} \gamma_{lk}^{-} + w_{lk} \gamma_{lk}^{+}$, $\gamma_{lk}^{-} = I(\gamma_{lk} = -1)$ and $\gamma_{lk}^{+} = I(\gamma_{lk} = +1)$. From (3),

$$\log\{p(y_{il} = y)\} = c + \begin{cases} 
\alpha_{i} \delta_{l}^{-} + \eta_{i}^{-} & \text{for } y = -1 \\
\alpha_{i} \delta_{l}^{+} + \eta_{i}^{+} & \text{for } y = +1 \\
0 & \text{for } y = 0
\end{cases}$$

(4)

where $\delta_{l} = (\delta_{1l}, \ldots, \delta_{Kl})^{T}$, $\delta_{l}^{-} = \delta_{l} \circ I(\gamma_{l} = -1)$, $\delta_{l}^{+} = \delta_{l} \circ I(\gamma_{l} = +1)$ with element-wise multiplication $\circ$, and $c = -\log\{ \exp(\alpha_{i} \delta_{l}^{-} + \eta_{i}^{-}) + \exp(\alpha_{i} \delta_{l}^{+} + \eta_{i}^{+}) + 1 \}$. Figure 3 illustrates the factorization. The matrix $A$ describes the PD relationships. $D^{-}$ and $D^{+}$ characterize the SD relationships where $D^{-} = D \circ I(C = -1)$ and $D^{+} = D \circ I(C = +1)$. The number of diseases is less than the number of patients and the number of symptoms.

**Edge-labeled random networks.** DFA can be also interpreted as inference for a random network with labeled edges. The observed categorical matrix $Y$ is treated as a categorical adjacency matrix which encodes a bipartite random network. Patients form one set of nodes and symptoms form another set. The edge labels correspond to the categories in $Y$. See the bipartite network on the upper portion of Figure 4 where the two labels ($+1$, $-1$) are respectively represented by arrow heads and flat bars. DFA assumes that the observed bipartite graph is generated from a latent tripartite graph (given in the lower portion of Figure 4). Inference under the DFA model reverses the data generation process. The tripartite graph introduces an additional set of (latent) nodes corresponding to diseases. The edges between patients (symptom) and diseases indicate PD (SD) relationships. Prior PD and SD knowledge is represented by fixing the corresponding edges (solid lines in Figure...
4 Posterior Inference

The model described in Section 2 is parameterized by \( \theta = \{ A, B, C, \{ W_j, \zeta_j \}_{j=1}^p, \{ W^+_l, W^-_l, \eta_l^-, \eta_l^+ \}_{l=1}^q \} \).

Posterior inference is carried out by Markov chain Monte Carlo (MCMC) posterior simulation. All parameters except \( A \) can be updated with simple Metropolis-Hasting transition probabilities. Sampling \( A \) is slightly more complicated because the dimension of the parameter space can change. We therefore provide details of the transition probability for updating \( A \). In the implementation, for easier bookkeeping, we set a large upper bound \( K_v = 50 \) for the number of latent diseases in the implementation; it is never reached during the course of the MCMC.

MCMC

Initialize \( \theta^{(0)} \). For \( t = 0, \ldots, T - 1 \), do
Figure 4: Edge-labeled random networks as an alternative representation of the DFA in Figure 2. The observed bipartite graph (top) is assumed to be generated by the latent tripartite graph (bottom). DFA addresses the inverse problem. Circles are patients, squares are symptoms and triangles are diseases whose colors have the same interpretations as in Figure 2. Dashed lines are latent whereas solid lines are known/observed. An undirected edge connecting patients and diseases is binary. Edges with arrow head or flat bars represent different types of SD relationships.

(1) Update $A^{(t+1)}$. We scan through each row, $i = 1, \ldots, n$, of $A = A^{(t)}$.

(1a) Update existing diseases $k = 1, \ldots, K$. If $\alpha_{-i,k} = 0$, drop disease $k$ (see Step 1b, below); otherwise, sample $\alpha_{ik}$ from the full conditional distribution,

$$p(\alpha_{ik} = 1 \mid \cdot) \propto p(\alpha_{ik} = 1 \mid \alpha_{-i,k}) p(z_{i} \mid \alpha_{i}, B, \{W_{j}, \zeta_{j}\}_{j=1}^{p}) p(y_{i} \mid \alpha_{i}, C, \{W_{l}^{-}, W_{l}^{+}, \eta_{l}^{--}, \eta_{l}^{++}\}_{l=1}^{q})$$

$$= \frac{r_{-i,k}}{n} \prod_{j=1}^{p} p(z_{ij} \mid \alpha_{i}, \beta_{j}, W_{j}, \zeta_{j}) \prod_{l=1}^{q} p(y_{il} \mid \alpha_{i}, \gamma_{l}, W_{l}^{-}, W_{l}^{+}, \eta_{l}^{--}, \eta_{l}^{++})$$

where $p(z_{ij} \mid \alpha_{i}, \beta_{j}, W_{j}, \zeta_{j})$ and $p(y_{ik} \mid \alpha_{i}, \gamma_{l}, W_{l}^{-}, W_{l}^{+}, \eta_{l}^{--}, \eta_{l}^{++})$ are given in (2) and (3).
(1b) If $K < K_v$, propose new diseases. The proposed new diseases are unique to patient $i$ only, i.e., $\alpha_{ik} = 1$ and $\alpha_{-i,k} = 0$. We first draw $k^* \sim \text{Poi}(m/n)$. If $k^* = 0$, go to the next step. Otherwise, we propose a set of new disease-specific parameters $\beta_k^* = (\beta_{1k}^*, \ldots, \beta_{pk}^*)^T$, $\gamma_k^* = (\gamma_{1k}^*, \ldots, \gamma_{qk}^*)^T$, $\{w_{jk}\}_{j=1}^p$, $\{w_{ik}^{-*}, w_{ik}^{+*}\}_{l=1}^q$ for $k = K + 1, \ldots, K + k^*$ from their respective prior distributions. Let $\text{Poi}(\cdot | \lambda)$ denote the Poisson probability mass function with rate $\lambda$. We accept the new disease(s) and the disease-specific parameters, with probability

$$
\min \left\{ 1, \frac{\prod_{j=1}^p p(z_{ij} \mid \alpha_i^*, \beta_j^*, W_j^*, \zeta_j)}{\prod_{j=1}^p p(z_{ij} \mid \alpha_i^{(t)}, \beta_j^{(t)}, W_j^{(t)}, \zeta_j)} \frac{\prod_{l=1}^q p(y_{il} \mid \alpha_i^*, \gamma_l^*, W_i^{-*}, W_i^{+*}, \eta_l^-, \eta_l^+)}{\prod_{l=1}^q p(y_{il} \mid \alpha_i^{(t)}, \gamma_l^{(t)}, W_i^{-{(t)}}, W_i^{+{(t)}}, \eta_l^-, \eta_l^+)} \right\},
$$

where $\alpha_i^* = (\alpha_i^{(t)}, 1, \ldots, 1)$, $\beta_j^* = (\beta_j^{(t)}T, \beta_{j,K+1}^*, \ldots, \beta_{j,K+k^*}^*)^T$, $\gamma_l^* = (\gamma_l^{(t)}T, \gamma_{l,K+1}^*, \ldots, \gamma_{l,K+k^*}^*)^T$, $W_j^* = \text{blkdiag}(W_j^{(t)}, w_{j,K+1}^*, \ldots, w_{j,K+k^*}^*)$, $W_i^{-*} = \text{blkdiag}(W_i^{-{(t)}}, w_{i,K+1}^*, \ldots, w_{i,K+k^*}^*)$ and $W_i^{+*} = \text{blkdiag}(W_i^{+{(t)}}, w_{i,K+1}^*, \ldots, w_{i,K+k^*}^*)$. Note that the acceptance probability only involves the likelihood ratio because prior and proposal probabilities are canceled out. If the new disease is accepted, we increase $K$ by $k^*$.

(2) Update all other parameters in $\theta^{(t+1)}$ using Metropolis-Hasting transition probabilities.

To summarize the posterior distribution based on the the MCMC simulation output, we proceed by first calculating the maximum a posteriori (MAP) estimate $\hat{K}$ from the marginal posterior distribution of $K$. Conditional on $\hat{K}$, we find the least squares estimator $\hat{A}$ by the following procedure (Dahl, 2006; Lee et al., 2015). For any two binary matrices $A, A' \in \{0, 1\}^{n \times \hat{K}}$, we define a distance $d(A, A') = \min_{\pi} \mathcal{H}(A, \pi(A'))$ where $\pi(A')$ denotes a permutation of the columns of $A'$ and $\mathcal{H}(\cdot, \cdot)$ is the Hamming distance of two binary matrices. A point estimate $\hat{A}$ is then obtained as

$$
\hat{A} = \arg \min_{A'} \int d(A, A') dp(A | Z, Y, \hat{K}).
$$

Both, the integral as well as the optimization can be approximated using the available Monte Carlo MCMC samples, by carrying out the minimization over $A' \in \{A^{(t)}; t = 1, \ldots, T\}$ and by evaluating the integral as Monte Carlo average. The posterior point estimators of other parameters in $\theta$ are obtained as posterior means conditional on $\hat{A}$. We evaluate the posterior
means using the posterior Monte Carlo samples.

5 Simulation Study

We consider two simulation scenarios. In both scenarios, we generate the patient-disease matrix $A$ from an IBP($m$) model with $m = 1$ and sample size $n = 300$. The resulting matrix $A$ has $K = 6$ columns and $n = 300$ rows, displayed in Figure 5(b). Given $K = 6$, we generate a binary symptom-disease matrix $B \in \{0, 1\}^{p \times K}$ and a categorical symptom-disease matrix $C \in \{-1, 0, 1\}^{q \times K}$ with $p = q = 24$ in the following manner. We first set $\beta_{jk} = 1$ for $k = 1, \ldots, 6$ and $j = 4(k - 1) + 1, \ldots, 4k$, $\gamma_{lk} = (-1)^k$ for $k = 1, \ldots, 6$ and $l = 4(k - 1) + 1, \ldots, 4k$. We then randomly change 10% of the zero entries in $B$ to 1 and 10% of the zero entries in $C$ to +1 or -1. The resulting matrices $B$ and $C$ are shown in Figures 5(c) and 5(d).

In Scenario I, the observations $Z$ and $Y$ are generated from the sampling model i.e. equations (2) and (3). To mimic the Chinese EHR data, we assume that we have diagnoses for the first disease and that we know the related symptoms for the first disease. In the model, we therefore fix the first column of $A$, $B$ and $C$ to the truth. In addition, we assume that we have partial information about the second latent disease: the symptom-disease relationships are known, but no diagnostic information is available. Accordingly, we will fix the second columns of $B$ and $C$, but leave the second column of $A$ as unknown parameters. We ran the MCMC algorithm described in Section 4 for 5,000 iterations, which took < 1 minute on a desktop computer with a 3.5 GHz Intel Core i7 processor. The first half of the iterations are discarded as burn-in and posterior samples are retained at every 5th iteration afterwards.

Inference summaries are reported in Figure 5. Figure 5(a) shows the posterior distribution of the number $K$ of latent diseases. The posterior mode occurs at the true value $\hat{K} = 6$. Conditional on $\hat{K}$, the posterior point estimate $\hat{A}$ is displayed in Figure 5(e) with misallocation rate $\mathcal{H}(\hat{A}, A)/(n \cdot K) = 3\%$\footnote{The percentage is computed based on free parameters in $A$ only.} Conditional on $\hat{A}$, the point estimates $\hat{B}$ and $\hat{C}$ are provided in Figures 5(f) and 5(g). The similarity between the heatmaps of simulation truth and estimates indicates an overall good recovery of the signal. The error rates in estimating $B$ and $C$ are 0% and 2%, respectively. We repeat this simulation 50 times. In 88% of the
repeat simulations, we correctly identify the number $K$ of latent diseases; in the remaining 12%, it is overestimated by 1. When $K$ is correctly estimated, the average mis-allocation rate, error rates for $B$ and $C$ are 3%, 1% and 1% with standard deviation 0.5%, 1% and 1%, respectively.

In Scenario II, we use a different simulation truth and generate data $\{\tilde{Z}, \tilde{Y}\}$ from latent factor models

$$\tilde{Z} = \Phi A_z + E_z \quad \text{and} \quad \tilde{Y} = \Phi A_y + E_y$$

with latent factor matrix $\Phi = A \circ \tilde{A}$ and loading matrices $A_z = B \circ \tilde{A}_z, A_y = C \circ \tilde{A}_y$ where $A, B, C$ are the same as in Scenario I. Each element of $\tilde{A}_z, \tilde{A}_y$ are i.i.d. $Unif(0, 4)$ and the errors are i.i.d. standard normal. We then threshold $\tilde{Z}$ and $\tilde{Y}$ to obtain $Z$ and $Y$ at different levels $t > 0$:

$$z_{ij} = \begin{cases} 1 & \text{if } \tilde{z}_{ij} > t \\ 0 & \text{if } \tilde{z}_{ij} \leq t \end{cases} \quad \text{and} \quad y_{il} = \begin{cases} +1 & \text{if } \tilde{y}_{il} > t \\ 0 & \text{if } |\tilde{y}_{il}| \leq t \\ -1 & \text{if } \tilde{y}_{il} < -t \end{cases}.$$ 

We applied DFA to $\{Z, Y\}$ using different values of the threshold $t \in \{1, 2, 3, 4, 5\}$. Also, we include no known diseases. Performance deteriorates as $t$ grows. DFA tends to overestimate the number $\hat{K}$ of latent factors by 1 to 3 extra factors. After removing those extra factors, the mis-allocation rate $H(\hat{A}, A)/(n \cdot K)$ is between 9% and 19%. And the error rates for $\hat{B}$ and $\hat{C}$ are between 1% and 17%.

For comparison, we applied the sparse latent factor models (SLFM, Ročková & George 2016) to $\{\tilde{Z}, \tilde{Y}\}$. SLFM assumes a sparse loading matrix and unstructured latent factors. Therefore, we only report performance in recovering the sparse structure $B$ and $C$ of the loading matrices $A_z$ and $A_y$. The penalty parameter $\lambda_0$ of SLFM is chosen in an “oracle” way: we fit SLFM with a range of $\lambda_0$ values and select $\lambda_0$ that yields the best performance given the simulation truth. The resulting error rates in estimating $B$ and $C$ are 11% and 10%, comparable to those of DFA (keeping in mind the oracle choice of $\lambda_0$).

We repeat the experiment 50 times for $t = 3$. The estimated $\hat{K}$ across 50 simulations are plotted in Figure 5(h). We observe that DFA tends to overestimate $K$ when model is
Figure 5: Simulation truth and posterior estimates. Panels (a)-(g) are from Scenario I and Panel (h) is from Scenario II. In the heatmaps, green cells represent 1, white cells 0 and red cells -1.
misspecified. We report the performance of estimating $\hat{B}$ and $\hat{C}$ based on the best subset of columns. The mean (standard deviation) error rates in estimating $B$ and $C$ are 8% (4%) and 9% (3%), respectively. SLFM has slightly higher error rates 12% (1%) and 11% (1%).

6 Phenotype Discovery with EHR Data

6.1 Data and preprocessing

We implement latent disease mining for the EHR data introduced in Section 1.1. Using the reference range for each test item, we define a symptom if the value of an item falls beyond the reference range. Some symptoms are binary in nature, e.g. low density lipoprotein is clinically relevant only when it is higher than normal range. Other symptoms are inherently ternary, e.g. heart rate is symptomatic when its too high or too low. The 39 testing items are listed in Table 1 where we also indicate whether each item gives rise to a binary or ternary symptom.

We extract diagnostic codes for diabetes from the sections “medical history” and “other current diseases” in the physical examination form. A subject is considered as having diabetes if it is listed in either of the two sections. There are 36 patients diagnosed with diabetes. We fix the first column of $A$ in the PD model according to the diabetes diagnosis. Moreover, it is known that diabetes is clinically associated with high glucose level. We incorporate this prior information by fixing the corresponding entry in the first column of $C$ in the SD model.

There is additional prior knowledge about symptom-disease relationships. Creatinine, a waste product from muscle metabolism, is controlled by the kidneys to stay within a normal range. Creatinine has therefore been found to be a reliable indicator of kidney function. Elevated creatinine level suggests damaged kidney function or kidney disease. Blood urea nitrogen (BUN) level is another indicator of kidney function. Like creatinine, urea is also a metabolic byproduct which can build up if kidney function is impaired. We fix the two entries (corresponding to creatinine and BUN) of the second column of $C$ to 1 and the rest to 0. With this prior knowledge, we interpret the second latent disease as kidney disease. Likewise, it is known that elevated systolic blood pressure and diastolic blood pressure are indicators of
hypertension, and abnormal levels of total bilirubin (TB), aspartate aminotransferase (AST) and alanine aminotransferase (ALT) are indicators of liver diseases. We fix the corresponding entries of the third and fourth column of $B$ and $C$, and interpret the third latent disease as *hypertension* and the fourth latent disease as *liver disease*.

To comply with Chinese policy, we report inference for data preprocessed by a Generative Adversarial Network (GAN, Goodfellow et al. 2014), which replicates the distribution underlying the raw data. GAN is a machine learning algorithm which simultaneously trains a generative model and a discriminative model on a training dataset (in our case, the raw EHR dataset). The generative model simulates a hypothetical repetition of the training data, which is then combined with the original training data to form a merged data set. Meanwhile, the discriminative model tries to distinguish between original data and simulations in the merged data set. During training, the generative model uses gradient information from the discriminative model to produce better simulations. Training continues until the discriminative model can no longer distinguish. After training, the generative model can be used to generate an arbitrary number of simulations which are similar in distribution to the original dataset. Any statistical inference in the original data and the replicated data is identical to the extent to which it relies on low dimensional marginal distributions. In our case, we generate a simulated dataset of the same size as the raw EHR dataset. The same approach has been used in Ni et al. (2018); see the detailed procedure therein.

### 6.2 Results and interpretations

We ran the MCMC algorithm described in Section 4 for 50,000 iterations. The first half of the iterations are discarded as burn-in and posterior samples are retained at every 5th iteration thereafter. The posterior distribution for $K = 14$ and $15$ is $p(K = 14) = 0.69$ and $p(K = 15) = 0.24$, respectively, i.e., the maximum a posteriori (MAP) estimate is $\hat{K} = 14$. This includes the 4 *a priori* known diseases as well as 10 newly discovered latent diseases.

Conditional on $\hat{K}$, the posterior estimates of the PD and SD models are shown as a heatmap in Figure 7 with green, black and red cells representing 1, 0 and -1, respectively. The nature of the figure as a single heatmap with two blocks, for PD and SD, respectively, highlights again the nature of the model as a “double” feature allocation generating matching subsets of patients and symptoms. The model allocates both, patients and symptoms, to
latent diseases. As mentioned before, DFA can also be interpreted as a edge-labeled network.

We show the same results as in the heatmap as a bipartite graph in Figure 8(a). The full inferred model would be a tripartite network, as in the bottom portion of Figure 4. However, we omit the patient nodes in Figure 8(a), lest the figure would be overwhelmed by the patient nodes. Instead, we summarize the patient-disease relationships by specifying the font size of the disease node (triangle, purple font) proportional to the number of linked patients. Latent diseases are labeled by numbers and a priori known diseases are labeled by name. Symptoms are shown in black font, with lines showing the links to diseases. Dashed lines are symptom-disease relationships inferred from the data whereas solid lines are fixed by prior knowledge. Black lines indicate that symptoms are binary. Red (blue) lines indicate suppression (enhancement) relative to the normal range. Line widths are proportional to the posterior probabilities of edge inclusion.

We find 239 patients with impaired kidney function or kidney disease, 183 patients with hypertension and 93 patients with liver disease. The prevalence of kidney disease is slightly higher than the national average 16.9% (Zhang et al., 2012) probably because of the elderly patient population in this study. The hypertension prevalence is much lower than the national average 57.3% (Zhang et al., 2017). We suspect it is due to the public awareness of hypertension and wide availability of the treatment.

We identify additional 10 latent diseases with prevalence 493, 218, 192, 174, 114, 82, 64, 24, 15, 15 patients. Some of the latent diseases are quite interesting. Latent disease 1 is lipid disorder, associated with high total cholesterol (TC), triglycerides and low density lipoprotein (LDL). Cholesterol is an organic molecule carried by lipoproteins. LDL is one type of such lipoproteins, commonly referred to as “bad” cholesterol. At normal levels, TC and LDL are essential substances for the body. However, high levels of TC and LDL put patients at increased risk for developing heart disease and stroke. Triglycerides are a type of fat found in the blood which are produced by the body from excessive carbohydrates and fats. Like cholesterol, triglycerides is essential to life at normal levels. However, a high level is associated with a greater chance for heart disease.

Latent disease 3 is thrombocytopenia-like disease which causes low count of platelets, decreased plateletcrit (PCT) and coefficient of variation of platelet distribution (PDW-CV), and increased mean platelet volume (MPV). Patients with low platelets may not be able to
stop bleeding after injury. In more serious cases, patients may bleed internally which is a life-threatening condition.

**Latent disease 4** is a *polycythemia-like* disease, associated with elevated mean corpuscular hemoglobin concentration (MCHC), hemoglobin, erythrocytes and hematocrit (HCT). These symptoms match exactly the symptoms of polycythemia, a disease that gives rise to an increased level of circulating red blood cells in the bloodstream. Polycythemia can be caused intrinsically by abnormalities in red blood cell production or by external factors such as chronic heart diseases.

Interestingly, like latent disease 4, **latent disease 6** is also related to hemoglobin, erythrocytes and HCT. However, it causes decreased rather than elevated levels of these items; hence we refer to the disease as *anemia*.

**Latent disease 7** suggests *bacterial infection* with increased leukocytes, granulocytes (GRA) and heart rate, and decreased monocytes (MON) and lymphocytes (LYM). The immune system, specifically the bone marrow, produces more GRA and leukocytes to fight a bacterial infection. As a result, the relative abundance of MON and LYM decreases.

Relatedly, **latent disease 8** may be caused by *viral infection*. Virus can disrupt the work of bone marrow which leads to low levels of leukocytes and GRA, and high levels of LYM.

**Latent disease 9** is related to *allergy* with abnormal basophil, GRA and LYM.

**Latent disease 10** suggests that a small group of patients may have *malnutrition* which leads to low blood glucose and anemia-like symptoms such as low corpuscular volume (MCV) and corpuscular hemoglobin (MCH).

Such automatically generated inference on latent disease phenotypes is of high value for routine health exams. It helps the practitioners to focus resources on specific patients and suggests meaningful additional reports. Inference in the statistical model can of course not replace clinical judgment and needs further validation. But it can provide an important decision tool to prioritize resources, especially for areas with limited medical support such as the areas where our data are collected from.

We remark that although there are clear interpretations for most latent diseases found by DFA, latent diseases 2 and 5 can not easily be interpreted as single diseases. Latent disease 2 is associated with 12 symptoms, which is likely beyond the number of symptoms of a realistic disease. Most of the symptoms, such as platelets, leukocytes and lymphocytes, are
due to weak immune system. Considering the elderly population of this dataset, aging could be a reasonable cause. Additionally, decreased heart rate and low glucose level can also be explained by aging. Latent disease 5 is connected to elevated systolic blood pressure and glucose. While those two symptoms may not be simultaneously linked to the same disease, their co-appearance should not be too surprising because the co-existence of hypertension and diabetes (to which blood pressure and glucose are knowingly linked) is in fact quite common (De Boer et al., 2017).

6.3 Web application for a disease diagnosis recommendation system

A good user interface is critical to facilitate the implementation of the proposed approach in the decision process of healthcare providers, and for broad application. Using the R package shiny (Chang et al., 2015), we have created an interactive web application (available at [https://nystat3.shinyapps.io/Rshiny/](https://nystat3.shinyapps.io/Rshiny/)). The application displays disease diagnosis recommendations for de-identified patients selected by the user. We show the application for two patients in Figure 6.

6.4 Comparison

As a comparison with results under alternative approaches, we considered inference under SLFM, applied directly to the blood test results, without converting to binary or ternary symptoms using the reference range. The tuning parameter $\lambda_0$ was set to 1 to approximately match the number of latent diseases found by DFA. SLFM implements inference on sparse symptom-disease relationships as shown in Figure 8(b). Although there is no known truth for symptom-disease relationships, it is difficult to interpret certain links. While uric acid may play some roles in certain diseases, we do not expect it to be related to 5 out of 12 latent diseases. In addition, both, latent diseases 5 and 6 are related to platelets only, which should be collapsed into one disease. We also ran LSFM with larger $\lambda_0$ but found similar results. For example, when $\lambda_0 = 10$, SLFM identifies 20 latent diseases, 17 of which are associated with uric acid. These somewhat surprising results may be due to the assumption of normality and linearity of SLFM, and taking no advantage of prior information.
Disease Diagnosis Recommendation System with De-identified Electronic Health Records

Use the drop-down menu to select patient identification number.

Patient ID#
0003

For selected patient, the disease recommendations are given (in black) together with associated symptoms (in red). Disease font size is proportional its prevalence.

---

Disease Diagnosis Recommendation System with De-identified Electronic Health Records

Use the drop-down menu to select patient identification number.

Patient ID#
0822

For selected patient, the disease recommendations are given (in black) together with associated symptoms (in red). Disease font size is proportional its prevalence.

---

Figure 6: Two examples of the R Shiny web application. Each example is for one patient. The application interactively displays disease diagnosis recommendations for de-identified patients selected by the user.
As already briefly commented in Section 1, graphical models may be also employed for finding hidden structures of the symptoms. We ran the birth-death MCMC algorithm (available in R package `BDgraph` [Mohammadi & Wit 2015]) for 50,000 iterations to learn a Bayesian graphical model. A point estimate is shown in Figure 8(c). The symptoms that form the cliques of length greater than 3 are represented by circles in Figure 8(c). Some of the findings are consistent with those by the DFA. As an illustration, the clique of TC, triglycerides, LDL and HDL are very similar to latent disease 1 in Figure 8(a). However, although graphical models can find latent patterns that are not immediately obvious in the correlation structure (Figure 1(a)), like SLFM, it lacks inference on patient-disease relationships. Moreover, using cliques or other graph summaries is an arbitrary choice.

7 Discussion

We have developed a DFA model for discovery of latent diseases in EHR data. DFA can be equivalently viewed as categorical matrix factorization or as inference in edge-labeled random networks. In the EHR data analysis, it is important to include available diagnostic information and other prior knowledge, which greatly facilitates identification of latent diseases. We found the prevalence of damaged kidney function in our dataset is comparable with, but slightly higher than the regional average. We also found 10 latent diseases that are related to lipid disorder, thrombocytopenia, polycythemia, anemia, bacterial and viral infections, allergy and malnutrition. Finally, although the proposed DFA model is specifically designed for disease mining, it could find broader applications in various areas such as education and psychology (Chen et al., 2015, 2018).

Though EHR often involves analysis of big datasets, the scalability to large sample size is not the focus of this paper. We focus on inference for data from a more narrowly restricted subset equivalent to about a week worth of data. We have shown that such data already allows meaningful inference on latent diseases. If desired and reasonable, the same approach could of course be used for larger data sets, but would likely need to be combined with model extensions to allow for changes in the latent structure, i.e., disease patterns, across different towns, times and other important factors. Implementation would need computationally efficient algorithms for posterior inference such as consensus Monte Carlo (Minsker et al.)
Figure 7: EHR data. The heatmap on top shows the estimated patient-disease relationships \( \hat{A} \). The bottom part of the double heatmap shows the estimated symptom-disease relationships \( \hat{B}, \hat{C} \) with green, black and red cells representing 1, 0 and -1, respectively. The columns correspond to diseases and the rows are patients (top portion) and symptoms (bottom portion).
Figure 8: EHR data analysis. (a) Bipartite network for symptom-disease relationships from DFA. The diseases are represented by triangles with the font size proportional to its popularity (i.e. the number of patients having the disease). Latent diseases are represented by the numbers (10 latent diseases in total). The symptoms are given in black font and their links to each disease are represented by the lines. Dashed lines are symptom-disease relationships inferred from the data whereas solid lines are fixed by prior knowledge. Black lines indicate the symptoms are binary. Red (blue) lines indicate the disease causes the symptom to be lower (higher) than normal range. The line width is proportional to its posterior probability of inclusion. (b) Bipartite network for symptom-disease relationships from SLFM. Latent diseases are represented by triangles (8 latent diseases in total). (c) Network for symptom-disease relationships from BDgraph. There are 5 cliques of length greater than 3 in total. The symptoms that form those cliques are represented by circles.
Another common feature of EHR data is massive missingness. Moreover, the assumption of missing completely at random usually does not hold. However, the EHR dataset that we considered here did not suffer from this feature.

References

Barrios, Ernesto, Lijoi, Antonio, Nieto-Barajas, Luis E., & Prünster, Igor. 2013. Modeling with Normalized Random Measure Mixture Models. *Statist. Sci.*, **28**(3), 313–334.

Bhattacharya, Anirban, & Dunson, David B. 2011. Sparse Bayesian infinite factor models. *Biometrika*, 291–306.

Broderick, Tamara, Jordan, Michael I, & Pitman, Jim. 2013. Cluster and feature modeling from combinatorial stochastic processes. *Statistical Science*, 289–312.

Campbell, Trevor, Cai, Diana, & Broderick, Tamara. 2018. Exchangeable trait allocations. *Electronic Journal of Statistics*, **12**(2), 2290–2322.

Chang, Winston, Cheng, Joe, Allaire, JJ, Xie, Yihui, & McPherson, Jonathan. 2015. *shiny: Web Application Framework for R*. R package version 0.12.2.

Chen, Yinghan, Culpepper, Steven Andrew, Chen, Yuguo, & Douglas, Jeffrey. 2018. Bayesian estimation of the DINA Q matrix. *Psychometrika*, **83**(1), 89–108.

Chen, Yunxiao, Liu, Jingchen, Xu, Gongjun, & Ying, Zhiliang. 2015. Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, **110**(510), 850–866.

Dahl, David B. 2006. Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian Inference for Gene Expression and Proteomics*, 201–218.

De Boer, Ian H, Bangalore, Sripal, Benetos, Athanase, Davis, Andrew M, Michos, Erin D, Muntner, Paul, Rossing, Peter, Zoungas, Sophia, & Bakris, George. 2017. Diabetes and hypertension: a position statement by the American Diabetes Association. *Diabetes Care*, **40**(9), 1273–1284.
Dobra, Adrian, Lenkoski, Alex, & Rodriguez, Abel. 2011. Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, **106**(496), 1418–1433.

Favaro, Stefano, & Teh, Yee Whye. 2013. MCMC for Normalized Random Measure Mixture Models. *Statist. Sci.*, **28**(3), 335–359.

Ghahramani, Zoubin, & Griffiths, Thomas L. 2006. Infinite latent feature models and the Indian buffet process. *Pages 475–482 of: Advances in Neural Information Processing Systems*.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, & Bengio, Yoshua. 2014. Generative Adversarial Nets. *Pages 2672–2680 of: Advances in Neural Information Processing Systems*.

Green, Peter J., & Thomas, Alun. 2013. Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika*, **100**(1), 91.

Guo, Lei. 2013. *Bayesian Biclustering on Discrete Data: Variable Selection Methods*. Ph.D. thesis, Harvard University.

Halpern, Yoni, Horng, Steven, Choi, Youngduck, & Sontag, David. 2016. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association*, **23**(4), 731–740.

Hartigan, John A. 1972. Direct clustering of a data matrix. *Journal of the American Statistical Association*, **67**(337), 123–129.

Henderson, Jette, Ho, Joyce C, Kho, Abel N, Denny, Joshua C, Malin, Bradley A, Sun, Jimeng, & Ghosh, Joydeep. 2017. Granite: Diversified, sparse tensor factorization for electronic health record-based phenotyping. *Pages 214–223 of: Healthcare Informatics (ICHI), 2017 IEEE International Conference on*. IEEE.

Lau, John W, & Green, Peter J. 2007. Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, **16**(3), 526–558.

Lauritzen, Steffen L. 1996. *Graphical models*. Vol. 17. Oxford: Clarendon Press.
Lee, Juhee, Müller, Peter, Gulukota, Kamalakar, & Ji, Yuan. 2015. A Bayesian feature allocation model for tumor heterogeneity. *Ann. Appl. Stat.*, 9(2), 621–639.

Li, Tao. 2005. A general model for clustering binary data. Pages 188–197 of: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM.

Meeds, Edward, Ghahramani, Zoubin, Neal, Radford M, & Roweis, Sam T. 2007. Modeling dyadic data with binary latent factors. Pages 977–984 of: *Advances in Neural Information Processing Systems*.

Miettinen, Pauli, Mielikäinen, Taneli, Gionis, Aristides, Das, Gautam, & Mannila, Heikki. 2008. The discrete basis problem. *IEEE Transactions on Knowledge and Data Engineering*, 20(10), 1348–1362.

Miller, Jeffrey W., & Harrison, Matthew T. 2017. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 0(0), 1–17.

Minsker, Stanislav, Srivastava, Sanvesh, Lin, Lizhen, & Dunson, David. 2014. Scalable and robust Bayesian inference via the median posterior. Pages 1656–1664 of: *International conference on machine learning*.

Mohammadi, Abdolreza, & Wit, Ernst C. 2015. BDgraph: an R package for Bayesian structure learning in graphical models. *arXiv preprint arXiv:1501.05108*.

Ni, Yang, Müller, Peter, Diesendruck, Maurice, Williamson, Sinead, Zhu, Yitan, & Ji, Yuan. 2018. Scalable Bayesian Nonparametric Clustering and Classification. *arXiv preprint arXiv:1806.02670*.

Pontes, Beatriz, Giráldez, Raúl, & Aguilar-Ruiz, Jesús S. 2015. Biclustering on expression data: A review. *Journal of Biomedical Informatics*, 57, 163–180.

Richardson, Sylvia, & Green, Peter J. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4), 731–792.
Ročková, Veronika, & George, Edward I. 2016. Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516), 1608–1622.

Ross, MK, Wei, Wei, & Ohno-Machado, L. 2014. Big data and the electronic health record. *Yearbook of medical informatics*, 9(1), 97.

Rukat, Tammo, Holmes, Chris C, Titsias, Michalis K, & Yau, Christopher. 2017. Bayesian boolean matrix factorisation. *arXiv preprint arXiv:1702.06166*.

Scott, Steven L, Blocker, Alexander W, Bonassi, Fernando V, Chipman, Hugh A, George, Edward I, & McCulloch, Robert E. 2016. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2), 78–88.

Uitert, Miranda van, Meuleman, Wouter, & Wessels, Lodewyk. 2008. Biclustering sparse binary genomic data. *Journal of Computational Biology*, 15(10), 1329–1345.

Wood, F., Griffiths, T. L., & Ghahramani, Z. 2006. A non-parametric Bayesian method for inferring hidden causes. *In: Proceedings of the Conference on Uncertainty in Artificial Intelligence*, vol. 22.

Xu, Yanxun, Lee, Juhee, Yuan, Yuan, Mitra, Riten, Liang, Shoudan, Müller, Peter, & Ji, Yuan. 2013. Nonparametric Bayesian bi-clustering for next generation sequencing count data. *Bayesian Analysis (Online)*, 8(4), 759.

Zhang, Fu-Liang, Guo, Zhen-Ni, Xing, Ying-Qi, Wu, Yan-Hua, Liu, Hao-Yuan, & Yang, Yi. 2017. Hypertension prevalence, awareness, treatment, and control in northeast China: a population-based cross-sectional survey. *Journal of Human Hypertension*, 32(1), 54.

Zhang, Luxia, Wang, Fang, Wang, Li, Wang, Wenke, Liu, Bicheng, Liu, Jian, Chen, Menghua, He, Qiang, Liao, Yunhua, Yu, Xueqing, et al. 2012. Prevalence of chronic kidney disease in China: a cross-sectional survey. *The Lancet*, 379(9818), 815–822.
Zhang, Zhongyuan, Li, Tao, Ding, Chris, & Zhang, Xiangsun. 2007. Binary matrix factorization with applications. Pages 391–400 of: Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. IEEE.