Electronic Health Record Analysis via Deep Poisson Factor Models

Ricardo Henao R.HENAO@DUKE.EDU
Electrical and Computer Engineering Department
Duke University
Durham, NC 27708, USA

James T. Lu JAMES.LU@DUKE.EDU
School of Medicine
Electrical and Computer Engineering Department
Duke University
Durham, NC 27708, USA

Joseph E. Lucas JOE@STAT.DUKE.EDU
Electrical and Computer Engineering Department
Duke University
Durham, NC 27708, USA

Jeffrey Ferranti JEFFREY.FERRANTI@DM.DUKE.EDU
School of Medicine
Duke University
Durham, NC 27708, USA

Lawrence Carin LCARIN@DUKE.EDU
Electrical and Computer Engineering Department
Duke University
Durham, NC 27708, USA

Editor: ...

Abstract

Electronic Health Record (EHR) phenotyping utilizes patient data captured through normal medical practice, to identify features that may represent computational medical phenotypes. These features may be used to identify at-risk patients and improve prediction of patient morbidity and mortality. We present a novel deep multi-modality architecture for EHR analysis (applicable to joint analysis of *multiple* forms of EHR data), based on Poisson Factor Analysis (PFA) modules. Each modality, composed of observed counts, is represented as a Poisson distribution, parameterized in terms of hidden binary units. Information from different modalities is shared via a deep hierarchy of common hidden units. Activation of these binary units occurs with probability characterized as Bernoulli-Poisson link functions, instead of more traditional logistic link functions. In addition, we demonstrate that PFA modules can be adapted to discriminative modalities. To compute model parameters, we derive efficient Markov Chain Monte Carlo (MCMC) inference that scales efficiently, with significant computational gains when compared to related models based on logistic link functions. To explore the utility of these models, we apply them to a subset of patients from the Duke-Durham patient cohort. We identified a cohort of over 12,000 patients with Type 2 Diabetes Mellitus (T2DM) based on diagnosis codes and laboratory tests out of our patient population of over 240,000. Examining the common hidden units uniting the PFA modules, we identify patient features that represent medical concepts. Experiments indicate that our learned features are better able to predict mortality and morbidity than clinical features identified previously in a large-scale clinical trial.

Keywords: Deep learning, multi-modality learning, Poisson factor model, electronic health records, phenotyping.

1. Introduction

Electronic health records (EHR) are quickly becoming a primary depository of detailed patient health information. These data, if properly analyzed, have the potential to be a nidus for novel insights that may improve patient diagnosis, treatment and safety. In particular, there has been increasing focus on utilizing such data to rapidly identify disease cohorts or "phenotypes" that can be leveraged in clinical and epidemiological studies. However, EHR data, a by-product of the often messy day-to-day interactions of physicians and patients in primary care hospital and emergency room settings, are often challenging to manipulate and interpret without expert input.

Many of the initial EHR phenotyping methods in the literature (Hripcsak and Albers, 2013; Mareedu et al., 2009) relied and continue to rely explicitly on heuristics generated through the collaboration of physicians and informatics. These "computable" phenotypes identified clusters of patients that, for example, suffer from a particular ailment (Newton et al., 2013). Computable phenotypes are often structured similar to decision trees that utilize multiple modes of patients data captured by the EHR[1] to filter patient groups. These modes may include physician and nursing notes from prior encounters, procedure and diagnosis codes, laboratory results, medications, radiology and pathology. Alternatively, other methods have relied on physician-labeled case and control samples (Chen et al., 2013), to identify patient features that may represent a patient phenotype.

Computable phenotypes resemble an analysis that physicians intuitively perform while diagnosing patients. At a high level, physicians assign patients to a latent space of plausible disease phenotypes that inform diagnosis and treatment. This assignment is based on heterogeneous data from the patient interview and physical exam, in combination with other data such as radiology reports, laboratory results and prior medication and medical history. For example, a young child who presents with multiple respiratory infections at an early age increases the probability of a cystic fibrosis phenotype, and thus may be a candidate for associated genetic testing.

Despite their success in (*i*) advancing medical record data mining across large medical institutions and (*ii*) genotype/phenotypes studies, efforts to develop computable phenotypes are nonetheless iterative, manual and difficult to scale. Further there appears to be widespread variability in disease definitions for even putatively "well-defined" diseases. In Richesson et al. (2013) it was shown that even for phenotypes where there is widespread agreement on the disease definition, such as Type 2 Diabetes Mellitus, definitions by different clinical groups captured different patient populations.

A complementary approach to modeling patient phenotypes from EHR data relies on utilizing unsupervised models. These computational phenotypes have the ability to identify

[1] See `https://phekb.org/`.

not only feature sets that represent known medical concepts, but they may also discover feature sets that may represent novel phenotypes that are: (*i*) subtypes of and/or (*ii*) run counter to clinically intuited groups. Applied to health-system data and CMS (Centers for Medicare & Medicaid Services) claims data, it has been demonstrated that sparse tensor factorization of multimodal patient data, transformed into count data, generates concise sets of sparse factors that are recognizable by medical professionals (Ho et al., 2014a,b). Patients can then be treated as a weighted composites of such factors.

While these automated models are efficient at extracting phenotype data and reducing manual input, they have several limitations (Chen et al., 2013; Ho et al., 2014a). Current models are unable to capture correlation both between and within data modes. For example, tensor factorization requires the presence of all modes of patient data within a limited time window to capture the patient-physician interaction. As the number of modes increase, the probability of all modes of data being captured within a limited time window decreases. This prevents leveraging subsets of data modes from (often) limited patient interactions with care givers. Meanwhile, models that concatenate multiple data modes, or evaluate each mode separately, lose correlation between data types. Additionally, current models do not allow one to integrate classification in a straightforward manner. Rather, prediction is conducted in a step-wise manner relying on defining factors first and then entering them into a classification procedure. Current models also often only incorporate a single layer of information, depriving the model of potentially rich higher-level correlation structure within and between modes.

Deep models, understood as multilayer modular networks, have recently generated significant interest from the machine learning community, in part because of their ability to obtain state-of-the-art performance in a wide variety of modalities. Commonly used modules include, but are not limited to, Restricted Boltzmann Machines (RBMs) (Hinton, 2002), Sigmoid Belief Networks (SBNs) (Neal, 1992), convolutional networks (LeCun et al., 1998), feedforward neural networks, and Dirichlet Processes (DPs) (Blei et al., 2004). Deep models are often employed in topic modeling, modeling data characterized by vectors of word counts. As discussed below, EHR data may often be expressed in terms of counts of entities (e.g., counts of types of medications or tests, generalizing the concept of words). Topic models are therefore of interest for EHR data. Examples of deep topic models, composed of DP modules, include the nested Chinese Restaurant Process (nCRP) (Blei et al., 2004), the hierarchical DP (HDP) (Teh et al., 2006), and the nested HDP (nHDP) (Paisley et al., 2015). Alternatively, topic models built using modules other than DPs have been proposed recently, for instance the Replicated Softmax Model (RSM) (Hinton and Salakhutdinov, 2009) based on RBMs, the Neural Autoregressive Density Estimator (NADE) (Larochelle and Lauly, 2012) based on neural networks, the Over-replicated Softmax Model (OSM) (Srivastava et al., 2013) based on DBMs, and Deep Poisson Factor Analysis (DPFA) (Gan et al., 2015a) based on SBNs.

DP-based models have attractive characteristics from the standpoint of interpretability, in the sense that their generative mechanism is parameterized in terms of *distributions over topics*, with each topic characterized by a *distribution over words*. Alternatively, non-DP-based models, in which modules are parameterized by a deep hierarchy of *binary units* (Hinton and Salakhutdinov, 2009; Larochelle and Lauly, 2012; Srivastava et al., 2013), do not have parameters that are as readily interpretable in terms of topics of this type, although model performance is often excellent. The DPFA model in Gan et al. (2015a) is one of the first representations that characterizes documents based on distributions over topics and words, while simultaneously employing a deep architecture based on binary units. Specifically, Gan et al. (2015a) integrates the capabilities of Poisson Factor Analysis (PFA) (Zhou et al., 2012) with a deep architecture composed of SBNs (Gan et al., 2015b). PFA is a nonnegative matrix factorization framework closely related to DP-based models. Results in Gan et al. (2015a) show that DPFA outperforms other well-known deep topic models.

Building on the success of DPFA, this paper proposes a new deep multi-modality architecture for topic modeling, based entirely on PFA modules. Our model merges two key aspects of DP and non-DP-based architectures, namely: (*i*) its nonnegative formulation relies on Dirichlet distributions, and is thus readily interpretable throughout all its layers, not just at the base layer as in DPFA (Gan et al., 2015a); (*ii*) it adopts the rationale of traditional non-DP-based models such as DBNs and DBMs, by connecting different modalities and layers via binary units, to enable learning of high-order statistics and structured correlations within and across modalities. The probability of a binary unit being on is controlled by a Bernoulli-Poisson link (Zhou, 2015) (rather than a logistic link, as in the SBN), allowing repeated application of PFA modules at all layers of the deep architecture.

The main contributions of this paper are as follows. (*i*) We develop a novel deep architecture for topic models based entirely on PFA modules. (*ii*) The model has inherent shrinkage in all its layers, thanks to the DP-like formulation of PFA. This is unlike DPFA, which is based on SBNs. (*iii*) The proposed model yields greatly improved mixing, compared to DPFA which requires sequential updates for its binary units; in our formulation these are updated in block. (*iv*) The proposed approach provides the ability to build deep multi-modality architectures and discriminative topic models with PFA modules. (*v*) We develop an efficient MCMC inference procedure that scales as a function of the number of *non-zeros* in the data and binary units. In contrast, models based on RBMs and SBNs scale with the size of the data and binary units. Finally, (*vi*) we demonstrate the applicability of this framework to the analysis of EHR data, with an associated interpretation of the inferred data features (topics and meta-topics, as detailed below).

2. Motivating Data

We utilize three modes of data: self-reported medication usage, laboratory tests, and diagnosis and procedure codes. Count matrices for each mode for each patient were extracted from a Duke University 5-year dataset. Specifically, we consider electronic health data generated from 2007 to 2011 in the care of Durham County residents within the Duke University Health System (DUHS), including three hospitals and an extensive network of outpatient clinics. This dataset includes over 240,000 patients with over 4.4 million patient visits.

2.1 Data Reconciliation

Patient data originated from the various hospitals and outpatient clinics of DUHS. As names for medications, laboratory tests and diagnosis and procedure codes are uniquely named at each facility, the data must first be reconciled to a common data dictionary.

Our dataset included 39,429 medication names. These names, which included both brand and generic names at various dosages and formulations, were mapped to their pharmaceutical active ingredients (AI) using a custom Python script that leveraged the RxNorm API[2]. RxNorm is a depository of medication information maintained by the National Library of Medicine and includes trade names, brand names, dosage information and active ingredients (Nelson et al., 2011). Compound medications that include multiple active ingredients incremented counts for all AI in that medication. We discovered 1,694 unique AI in our dataset.

The data also include 4,391 types of laboratory tests, mapped to the Logical Observation Identifiers Names and Codes (LOINC) Ontology (Vreeman et al., 2015). The LOINC standard is common terminology for laboratory and clinical observations maintained by the Regenstrief Institute[3]. Mappings to the LOINC database were performed with the RELMA tool[4]. Each suggested mapping was reviewed by a physician to ensure that appropriate test and measurement units were aligned. Counts for patient laboratory tests reflect the number of times an AI appears in a patients record. We discovered 1,869 unique LOINC tests in the data.

Lastly, the data include 21,305 diagnosis and procedure codes. These were mapped using their unique ICD-9 (International Statistical Classification of Diseases and Related Health Problem) and CPT (Current Procedural Terminology) identifiers. ICD codes are the international diagnostic coding system, and are maintained by the World Health Organization[5]. CPT procedure codes are maintained by the American Medical Association and are designed to ease uniform communication performed medical services[6]. We identified 21,013 unique ICD-9 and CPT codes in the dataset.

2.2 Cohort and Count Matrix Generation

To narrow our analysis, we focused on a cohort of Type-2 Diabetes Mellitus (T2DM) patients, using previous phenotype criteria for T2DM (Richesson et al., 2013). T2DM is a chronic disease with high disease and treatment costs. Patients with diabetes are at increased risk of complications such as coronary heart disease (CHD), acute myocardial infarction (AMI), cerebral vascular disease (CVD), chronic renal failure (CRF), and amputation (American Diabetes Association, 2014). Prediction of these outcomes is important for communicating prognosis and targeting treatment to the high-risk patients.

We identified 16,756 patients in the dataset, by filtering for the following criteria: (*i*) at least two counts of 250.xx ICD-9 codes, (*ii*) at least one laboratory measurement of hemoglobin A1c (HgbA1C) greater than 4.5%, and (*iii*) a medication record including at least one of the following Type-2 diabetes medications: insulin, metformin, sufonylurea, or sitagliptin. We generated counts for each data mode by mapping each patient's records to the common data elements as described above. We then counted the total number of occurrences for each data element over a defined time window. In our initial experiment exploring the mapping of medical concepts to discovered factors, this time window represented two years of data. In our classification experiment, this time window was six months prior to the classification date.

[2] See `https://rxnav.nlm.nih.gov/RxNormAPIs.html`.

[3] See `https://loinc.org/`.

[4] See `https://loinc.org/downloads/relma`.

[5] See `http://www.who.int/classifications/icd/en/`.

[6] See `http://www.ama-assn.org/ama`.

2.3 Classification

2.3.1 UK Prospective Diabetes Study (UKPDS) Outcomes Model

Prediction equations to determine the risk of various complications in diabetes have been studied extensively (Wilson et al., 1998; Clarke et al., 2004; van Dieren et al., 2011). These risk estimates are helpful for identifying high-risk populations that may need closer clinical observation and higher intensity treatment (Simmons et al., 2009). Several equations are currently available to estimate CHD, AMI and CVD risk (Metcalf et al., 2008; Tao et al., 2013). UKPDS is a multicenter randomized trial involving 5,102 newly diagnosed patients with T2DM, recruited from 23 UK centers (King et al., 1999; Stevens et al., 2001); it has been utilized to generate outcome models for cardiovascular and cerebrovascular disease (Lu et al., 2012; Tao et al., 2013).

In the UKPDS model, the 1-year probability of CHD is:

$$\begin{aligned} p(\mathrm{CHD}) \propto\ & b_0 + b_1 * (\mathrm{Age} - 55) - b_2\mathrm{Female} - b_3\mathrm{AfroCaribbean} \\ & + b_4\mathrm{Smoking} + b_5(\mathrm{HgbA1c} - 6.72) + b_6(\mathrm{SPB} - 135.7)/10 \\ & + b_7(\log(\mathrm{TC/HDL}) - 1.59)\,, \end{aligned} \tag{1}$$

where mean total cholesterol (TC), mean high density lipoprotien (HDL), mean systolic blood pressure (SPB), and mean hemoglobin A1c (HgbA1c) are used, and $\{b_i\}_{i=0}^{7}$ are prespecified classification weights (Stevens et al., 2001).

While patient care has changed rapidly since this study was performed (the original patients were recruited and followed prospectively from 1977-1997), numerous studies have since explored its application in more recent clinical cohorts. These differences as well as regional variation in health care access and disease burden compelled us to estimate the UKPDS parameters in our cohort to improve its classification results for our patients.

2.3.2 Outcomes Identification

We generated training and test cohorts for our classification experiment in Section 4.3 by defining well-known T2DM disease morbidities with diagnosis and procedure codes (American Diabetes Association, 2014). For each patient we capture the date of the 13 outcomes in Table 1.

2.3.3 Generating Test and Train Cohorts

To generate training and test cohorts from our dataset, we selected a reference date that allowed us to (*i*) capture a large patient population with multiple patient encounters prior to the date, and (*ii*) evaluate encounters after the date for the existence of an outcome.

For the classification experiment we generated count matrices for each data mode by aggregating patient data for a six month period prior to the *patient visit*. We then determined if the patient had one or more of the above outcomes within 1 year of that visit. For the training cohort, the *patient visit* was defined as the encounter immediately prior to the date threshold of January 1, 2010. For our test set, we used a January 1, 2011 threshold date. We cleaned our data to remove any patients (*i*) that already had outcomes 6 months prior to the *patient visit* and (*ii*) removed any codes with less than 10 observations over the entire cohort. Lastly we removed from the test set any patients that were in the original

Outcome	ICD-9 codes	CPT codes
Acute Myocardial Infarction	410.*	—
Amputation	84.1*	—
Cardiac Catheterization	—	37.2*
Coronary Artery Disease	411.*, 413* and 414*	45.8*
Depression	293.*, 296.*, 300.4 and 311.*	—
Heart Failure	428.*	—
Kidney Disease	585.*, 249.* and 250.4*	56.1*
Neurological Diseases	249.6* and 250.6*	—
Obesity	278.*	85.*
Ophthalmic Disease	249.5* and 250.5*	—
Stroke	346.6*, 430.*, 431.*, 432.*, 433.*, 434.* and 435*	—
Unstable Angina	411.1	—
Death	date of death in the medical record	

Table 1: Definition of the 13 T2DM related outcomes for multi-label classification experiment in Section 3.5.

test set and did not have any additional visits since the training set threshold date. We also removed any individuals who did not have laboratory or vitals data in the prior 2 years, preventing us from computing a UKPDS risk score.

3. Model

3.1 Poisson factor analysis as a module

Assume $\mathbf{x}_n$ is an M-dimensional vector containing counts of M different entities (e.g., words in documents), for the n-th of N data vectors. We impose the model

$$\mathbf{x}_n \sim \text{Poisson}\left(\mathbf{\Psi}(\boldsymbol{\theta}_n \circ \mathbf{h}_n)\right) \,, \tag{2}$$

where $\mathbf{\Psi} \in \mathbb{R}_+^{M\times K}$ is the factor loadings matrix with K factors, $\boldsymbol{\theta}_n \in \mathbb{R}_+^K$ are factor intensities, $\mathbf{h}_n \in \{0,1\}^K$ is a vector of binary units indicating which factors are active for observation n, and $\circ$ represents the element-wise (Hadamard) product. The representation in (2) may be expressed as

$$x_{mn} = \textstyle\sum_{k=1}^K x_{mkn} \,, \qquad x_{mkn} \sim \text{ Poisson}(\lambda_{mkn}) \,, \qquad \lambda_{mkn} = \psi_{mk}\theta_{kn}h_{kn} \tag{3}$$

where $\boldsymbol{\psi}_k$ is column k of $\mathbf{\Psi}$, ψ_{mk} is component m of $\boldsymbol{\psi}_k$, x_{mn} is component m of $\mathbf{x}_n$, θ_{kn} is component k of $\boldsymbol{\theta}_n$, and h_{kn} is component k of $\mathbf{h}_n$. In (3) we have used the additive property of the Poisson distribution to decompose the m-th observed count of $\mathbf{x}_n$ as K latent counts, $\{x_{mkn}\}_{k=1}^K$.

One possible prior specification for this model, recently introduced in Zhou et al. (2012), is

$$\boldsymbol{\psi}_k \sim \text{ Dirichlet}(\eta \mathbf{1}_M) \,, \qquad \theta_{kn} \sim \text{ Gamma}(r_k, (1-b)b^{-1}) \,, \qquad h_{kn} \sim \text{ Bernoulli}(\pi_{kn}) \tag{4}$$

where $\mathbf{1}_M$ is an M-dimensional vector of all-ones. Furthermore, for simplicity we let $\eta = 1/K$, $b = 0.5$ and $r_k \sim \text{Gamma}(1,1)$. Prior distributions for η and b that result in closed

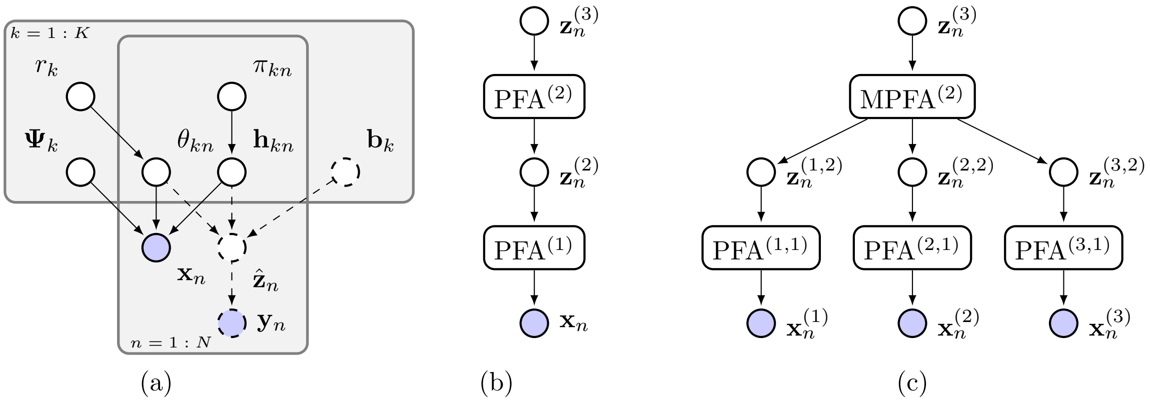


Figure 1: Graphical models. (a) Poisson Factor analysis module in (4). Nodes ($\mathbf{b}_k$, $\hat{\mathbf{z}}_n$ and $\mathbf{y}_n$) and edges drawn with dashed lines correspond to the discriminative PFA described in Section 3.5. (b) Deep Poisson factor model in (6). (c) Deep Multi-task Poisson factor model in (7). Filled and empty nodes represent observed and latent variables, respectively.

form conditionals exist, and can be used if desired; see for instance Escobar and West (1995) for η, and Zhou and Carin (2015) for b.

There is one parameter in (4) for which we have not specified a prior distribution, specifically $\mathbb{E}[p(h_{kn} = 1)] = \pi_{kn}$. In Zhou et al. (2012), h_{kn} is provided with a beta-Bernoulli process prior by letting $\pi_{kn} = \pi_k \sim \text{Beta}(c\epsilon, c(1-\epsilon))$, where usually $c = 1$ and $\epsilon = 1/K$, meaning that each of the N data vectors has on average the same probability of seeing a particular topic as active. It further assumes topics are independent of each other. These two assumptions are restrictive because: (i) in practice, the N data vectors often belong to a heterogeneous population (e.g., patients); letting the data vectors have individual topic activation probabilities allows the model to better accommodate heterogeneity in the data. (ii) Some topics are likely to co-occur systematically, so being able to harness such correlation structures can improve the ability of the model for fitting the data.

The hierarchical model in (2)-(4) is denoted $\mathbf{x}_n \sim \text{PFA}(\mathbf{\Psi}, \boldsymbol{\theta}_n, \mathbf{h}_n; \eta, r_k, b)$, short for Poisson Factor Analysis (PFA), with graphical model representation shown in Figure 1(a). The model in (2)-(4) is closely related to other widely known topic model approaches, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), HDP (Teh et al., 2006) and Focused Topic Modeling (FTM) (Williamson et al., 2010). Connections between these models are discussed in Section 3.7.

3.2 Deep representations with PFA modules

Several models have been proposed recently to address the limitations described in the previous section (Blei et al., 2004; Blei and Lafferty, 2007; Gan et al., 2015a; Teh et al., 2006). In particular, Gan et al. (2015a) proposed using multilayer SBNs (Neal, 1992) to impose correlation structure across topics, while providing each data vector with the ability to control its topic activation probabilities, without the need of a beta-Bernoulli process (Zhou et al., 2012). Here we follow the same rationale as Gan et al. (2015a), but without

SBNs. We start by noting that for a binary vector $\mathbf{h}_n$ with elements h_{kn}, we can write

$$h_{kn} = 1(z_{kn} \geq 1), \qquad z_{kn} \sim \text{Poisson}(\tilde{\lambda}_{kn}) \,, \tag{5}$$

where z_{kn} is a latent count for variable h_{kn}, parameterized by a Poisson distribution with rate $\tilde{\lambda}_{kn}$. The function $1(\cdot)$ is defined as $1(\cdot) = 1$ if the argument is true, and $1(\cdot) = 0$ otherwise. The model in (5), recently proposed in Zhou (2015), is known as the Bernoulli-Poisson Link (BPL) and is denoted $\mathbf{h}_n \sim \text{BPL}(\tilde{\boldsymbol{\lambda}}_n)$, for $\tilde{\boldsymbol{\lambda}}_n \in \mathbb{R}_+^K$. After marginalizing out the latent count z_{kn} (Zhou, 2015), the model in (5) has the interesting property that $p(h_{kn} = 1) = \text{Bernoulli}(\pi_{kn})$, where $\pi_{kn} = 1 - \exp(-\tilde{\lambda}_{kn})$. In order to sample h_{kn} we do not need to instantiate latent count z_{kn} but the rate of its underlying distribution $\tilde{\lambda}_{kn}$. Hence, rather than using the logistic function to represent binary unit probabilities as in SBNs, we employ $\pi_{kn} = 1 - \exp(-\tilde{\lambda}_{kn})$.

The binary distribution based on $p(h = 1) = \text{Bernoulli}(1 - \exp(-\tilde{\lambda}))$ is reminiscent of the complementary log-log link function (Piegorsch, 1992; Collett, 2002), where $\tilde{\lambda} = \exp(-u)$ and $u \in \mathbb{R}$. Unlike the logistic function, that is symmetric around the origin, $u = 0$ for $p(h = 1) = \text{Bernoulli}(1/(1 + \exp(-u)))$, the complementary log-log link is asymmetric, making it appropriate for imbalanced modalities, where the proportion of zeros is large. In our setting, this behavior is ideal because it encourages sparsity, in that it supports the assumption that a given data vector (patient) is explained by a small subset of topics selected via binary units, $\mathbf{h}_n$.

In (3) and (5) we have represented the Poisson rates as λ_{mkn} and $\tilde{\lambda}_{kn}$, respectively, to distinguish between the two. However, the fact that the count vector in (4) and the binary variable in (5) are both represented in terms of Poisson distributions suggests the following deep model, based on PFA modules

$$\begin{aligned}
\mathbf{x}_n &\sim \text{PFA}\left(\boldsymbol{\Psi}^{(1)}, \boldsymbol{\theta}_n^{(1)}, \mathbf{h}_n^{(1)}; \eta^{(1)}, r_k^{(1)}, b^{(1)}\right), & \mathbf{h}_n^{(1)} &= \mathbf{1}\left(\mathbf{z}_n^{(2)}\right), \\
\mathbf{z}_n^{(2)} &\sim \text{PFA}\left(\boldsymbol{\Psi}^{(2)}, \boldsymbol{\theta}_n^{(2)}, \mathbf{h}_n^{(2)}; \eta^{(2)}, r_k^{(2)}, b^{(2)}\right), & &\vdots \\
&\vdots & \mathbf{h}_n^{(L-1)} &= \mathbf{1}\left(\mathbf{z}_n^{(L)}\right), \\
\mathbf{z}_n^{(L)} &\sim \text{PFA}\left(\boldsymbol{\Psi}^{(L)}, \boldsymbol{\theta}_n^{(L)}, \mathbf{h}_n^{(L)}; \eta^{(L)}, r_k^{(L)}, b^{(L)}\right), & \mathbf{h}_n^{(L)} &= \mathbf{1}\left(\mathbf{z}_n^{(L+1)}\right),
\end{aligned} \tag{6}$$

where L is the number of layers in the model, and $\mathbf{1}(\cdot)$ is a vector operation in which each component imposes the left operation in (5). In this Deep Poisson Factor Model (DPFM), shown as a graphical model in Figure 1(b), the binary units at layer $\ell \in \{1, \ldots, L\}$ are drawn $\mathbf{h}_n^{(\ell)} \sim \text{BPL}(\boldsymbol{\lambda}_n^{(\ell+1)})$, for $\boldsymbol{\lambda}_n^{(\ell)} = \boldsymbol{\Psi}^{(\ell)}(\boldsymbol{\theta}_n^{(\ell)} \circ \mathbf{h}_n^{(\ell)})$. The form of the model in (6) introduces latent variables $\{\mathbf{z}_n^{(\ell)}\}_{\ell=2}^{L+1}$ and the element-wise function $\mathbf{1}(\cdot)$, rather than explicitly drawing $\{\mathbf{h}_n^{(\ell)}\}_{\ell=1}^{L}$ from the BPL distribution. Concerning the top layer, we let $z_{kn}^{(L+1)} \sim \text{Poisson}(\lambda_k^{(L+1)})$ and $\lambda_k^{(L+1)} \sim \text{Gamma}(a_0, b_0)$.

3.3 Deep multi-modality representation with PFA modules

In a multi-modality setting, each individual is characterized by D different count vectors, each of which is characterized by a different vocabulary (different types of entities being counted, i.e., different modalities). Individual $n \in \{1,\dots,N\}$, data type $i \in \{1,\dots,D\}$ is denoted as $\mathbf{x}_n^{(i)}$, corresponding to an M_i-dimensional count vector. The dataset described in Section 2 is composed of $D = 3$ data types: medications, laboratory tests and codes. Since all D data types are composed of count vectors, we can in principle concatenate the D vectors for patient n into a long $\sum_i M_i$-dimensional vector, $[(\mathbf{x}_n^{(1)})^\top,\dots,(\mathbf{x}_n^{(D)})^\top]^\top$, that we can model with the DPFM in (6). Such an approach will allow us to learn about the correlation structure of the variables in the concatenated modalities, but it ignores the fact that due to context, each data type in general has its own correlation structure. Another simple approach consists of modeling each data type individually, again using (6); however, this fails to acknowledge that different modality types can be correlated, as they represent different contexts or "views" of a larger representational space. Motivated by the shortcoming of these two simplistic approaches, we modify the model in (6) to learn correlation structures of individual modalities, but at the same time to be able to share information across them to leverage their correlation structure. In particular, we propose a data-type-specific first layer and a deep architecture of shared PFA modules, formally written as

$$\begin{aligned}
\mathbf{x}_n^{(i)} &\sim \mathrm{PFA}^{(i,1)}\,, & \mathbf{h}_n^{(i,1)} &= \mathbf{1}\left(\mathbf{z}_n^{(i,2)}\right)\,, \qquad i = 1,\dots,D\,,\\
\mathbf{z}_n^{(1,2)},\dots,\mathbf{z}_n^{(D,2)} &\sim \mathrm{MPFA}^{(2)}\,, & &\vdots\\
&\vdots & \mathbf{h}_n^{(L-1)} &= \mathbf{1}\left(\mathbf{z}_n^{(L)}\right)\,,\\
\mathbf{z}_n^{(L)} &\sim \mathrm{PFA}^{(L)}\,, & \mathbf{h}_n^{(L)} &= \mathbf{1}\left(\mathbf{z}_n^{(L+1)}\right)\,,
\end{aligned} \tag{7}$$

where

$$\mathrm{PFA}^{(i,1)} \overset{\text{def}}{=} \mathrm{PFA}\left(\boldsymbol{\Psi}^{(i,1)},\boldsymbol{\theta}_n^{(i,1)},\mathbf{h}_n^{(i,1)};\eta^{(i,1)},r_k^{(i,1)},b^{(i,1)}\right)\,, \tag{8}$$

$$\mathrm{MPFA}^{(2)} \overset{\text{def}}{=} \prod_i^D \mathrm{PFA}\left(\boldsymbol{\Psi}^{(i,2)},\boldsymbol{\theta}_n^{(2)},\mathbf{h}_n^{(2)};\eta^{(2)},r_k^{(2)},b^{(2)}\right)\,, \tag{9}$$

$$\mathbf{z}_n^{(i,2)} \sim \mathrm{PFA}\left(\boldsymbol{\Psi}^{(i,2)},\boldsymbol{\theta}_n^{(2)},\mathbf{h}_n^{(2)};\eta^{(2)},r_k^{(2)},b^{(2)}\right)\,. \tag{10}$$

The first layer in (7) is composed of D independent PFA modules as in (8), with explicit hierarchical model in (4). The multi-modality PFA model, denoted MPFA in (9), is a PFA model in which each modality has an associated factor loadings matrix, $\boldsymbol{\Psi}^{(i,2)}$, but shared factor intensities, $\boldsymbol{\theta}_n^{(2)}$, binary units, $\mathbf{h}_k^{(2)}$ and parameters $\{\eta^{(2)},r_k^{(2)},b^{(2)}\}$. This means that modality-specific latent counts, $\mathbf{z}_n^{(i,2)}$, can be drawn from a PFA module restricted to $\boldsymbol{\Psi}^{(i,2)}$ as in (10). The architecture of the Deep Multi-modality Poisson Factor Model (DMPFM) in (7) is fully specified by $\{K_{(1,1)},\dots,K_{(D,1)},K_{(2)},\dots,K_{(L)}\}$ and L, where $K_{(i,1)}$ are modality-specific loadings sizes (number of topics), $K_{(k)}$ are modality-shared loadings sizes and L is the number of layers. For example, Figure 1 shows a graphical model representation for a specification with $D = 3$ and $L = 2$.

3.4 Model interpretation

Consider modality i in layer 1 of (7), from which $\mathbf{x}_n^{(i)}$ is drawn. Assuming $\mathbf{h}_n^{(i,1)}$ is known, this corresponds to a focused topic model (Williamson et al., 2010). The columns of $\mathbf{\Psi}^{(i,1)}$ correspond to modality-i topics, with the k-th column $\boldsymbol{\psi}_k^{(i,1)}$ defining the probability with which entities (e.g., medications) of modality i are manifested for topic k (each $\boldsymbol{\psi}_k^{(i,1)}$ is drawn from a Dirichlet distribution, as in (4)). Generalizing the notation from (3), $\boldsymbol{\lambda}_{kn}^{(i,1)} = \boldsymbol{\psi}_k^{(i,1)} \theta_{kn}^{(i,1)} h_{kn}^{(i,1)} \in \mathbb{R}_+^M$ is the rate vector associated with topic k, modality i and patient n, and it is active when $h_{kn}^{(i,1)} = 1$. The entity-count vector for patient n in modality i manifested from topic k is $\mathbf{x}_{kn}^{(i)} \sim \text{Poisson}(\boldsymbol{\lambda}_{kn}^{(i,1)})$, and $\mathbf{x}_n^{(i)} = \sum_{k=1}^{K_{(i,1)}} \mathbf{x}_{kn}^{(i)}$, where $K_{(i,1)}$ is the number of topics in the module. The columns of $\mathbf{\Psi}^{(i,1)}$ define correlation among the entities associated with the topics; for a given topic (column of $\mathbf{\Psi}^{(i,1)}$), some entities co-occur with high probability, and other entities are likely jointly absent.

We now consider a two-layer model, with $\mathbf{h}_n^{(2)}$ assumed known. To generate $\mathbf{h}_n^{(i,1)}$, we first draw $\mathbf{z}_n^{(i,2)}$, which, analogous to above, may be expressed as $\mathbf{z}_n^{(i,2)} = \sum_{k=1}^{K_2} \mathbf{z}_{kn}^{(i,2)}$, with $\mathbf{z}_{kn}^{(i,2)} \sim \text{Poisson}(\boldsymbol{\lambda}_{kn}^{(i,2)})$ and $\boldsymbol{\lambda}_{kn}^{(i,2)} = \boldsymbol{\psi}_k^{(i,2)} \theta_{kn}^{(2)} h_{kn}^{(2)}$. Note that factor intensities and binary units, respectively $\theta_{kn}^{(2)}$ and $h_{kn}^{(2)}$, are shared across the $i = 1, \ldots, D$ modalities. Column k of $\mathbf{\Psi}^{(i,2)}$ corresponds to a *meta-topic* specific to modality i, with $\boldsymbol{\psi}_k^{(i,2)}$ a $K_{(i,1)}$-dimensional probability vector, denoting the probability with which each of the modality-i layer-1 topics are "on" when layer-2 "meta-topic" k is on (i.e., when $h_{kn}^{(2)} = 1$). The columns of $\mathbf{\Psi}^{(i,2)}$ define correlation among the modality-i layer-1 topics; for a given layer-2 meta-topic (column of $\mathbf{\Psi}^{(i,2)}$), some layer-1 topics co-occur with high probability, and other layer-1 topics are likely jointly absent. Furthermore, columns of the concatenated meta-topic matrix, $[(\mathbf{\Psi}^{(1,2)})^\top \ \ldots \ (\mathbf{\Psi}^{(D,2)})^\top]^\top$, define correlation structure among all layer-1 topics at the same time.

As one moves up the hierarchy, to layers $\ell > 2$, the meta-topics become increasingly more abstract and sophisticated, manifested in terms of probabilisitic combinations of topics and meta-topics at the layers below. Because of the properties of the Dirichlet distribution, each column of a particular $\mathbf{\Psi}^{(\ell)}$ is encouraged to be sparse, implying that a column of $\mathbf{\Psi}^{(\ell)}$ encourages use of a small subset of columns of $\mathbf{\Psi}^{(\ell-1)}$, with this repeated all the way down to the data layer, and the topics reflected in the columns of $\mathbf{\Psi}^{(1)}$. This deep architecture imposes correlation across the layer-1 topics in all modalities, and it does it through use of PFA modules at all layers of the deep architecture, unlike Gan et al. (2015a) which uses an SBN for layers 2 through L, and a PFA at the bottom layer. In addition to the elegance of using a single class of modules at each layer, the proposed deep model has important computational benefits, as discussed in Section 3.6.

We emphasize that $\{\theta_{kn}^{(2)}, h_{kn}^{(2)}\}$ are shared across all D data types, or modalities. The hierarchy that resides above them is meant to model underlying latent correlations in aspects of disease and health. The underlying state of the patient is independent of the modality with which he/she is viewed. When $h_{kn}^{(2)} = 1$, the kth meta-topic of health/disease is "on" for patient n; $\boldsymbol{\psi}_k^{(i,2)}$ characterizes how meta-topic k impacts the presence/absence of each topic associated with modality i. The modality-dependence is manifested at the bottom of the deep model, near the data, and the deep architecture above it imposes statistical relationships among the meta-topics, and is meant to characterize latent health/disease.

3.5 PFA modules for multi-label classification

Assume there is a C-dimensional vector of binary labels $\mathbf{y}_n \in \{0,1\}^C$ associated with patient n (presence/absence of C maladies or illnesses). Provided that labels share the same covariates (patient n, $\mathbf{x}_n$) and are oftentimes correlated, it is reasonable to model all labels jointly as opposed to build individual models for each label. We seek to learn the model for mapping $\mathbf{x}_n \to \mathbf{y}_n$ simultaneously with learning the deep topic representation in Section (3.2). In fact, the mapping $\mathbf{x}_n \to \mathbf{y}_n$ is based on the deep generative process for $\mathbf{x}_n$ in (6). This means that we can leverage the correlation structure of count data vectors and labels at the same time. We represent each element of $\mathbf{y}_n$, y_{cn}, using (5). We impose the model

$$y_{cn} = 1(\hat{z}_{cn} \geq 1)\,, \qquad \hat{z}_{cn} \sim \text{Poisson}(\hat{\lambda}_{cn})\,, \tag{11}$$

where $\hat{\lambda}_{cn}$ is element c of $\hat{\boldsymbol{\lambda}}_n$. First considering the single-modality case, $\hat{\boldsymbol{\lambda}}_n = \mathbf{B}(\boldsymbol{\theta}_n^{(1)} \circ \mathbf{h}_n^{(1)})$ and $\mathbf{B} \in \mathbb{R}_+^{C\times K}$ is a matrix of nonnegative classification weights, with prior distribution $\mathbf{b}_k \sim \text{Dirichlet}(\zeta \mathbf{1}_C)$, where $\mathbf{b}_k$ is a column of $\mathbf{B}$. Here, we denote latent counts as $\hat{\mathbf{z}}_n = [\hat{z}_{1n}\ \ldots\ \hat{z}_{Cn}]^\top$ to differentiate them form those coming from the DPFM, denoted as $\mathbf{z}_n^{(\ell)}$ in (6). The matrix of classification weights, $\mathbf{B}$, in (11) serves two purposes: (i) learns the correlation structure of labels, since large entries in $\mathbf{b}_k$, say b_{ck} and $b_{c'k}$ indicate their corresponding labels, c and c' are proportionally correlated; and (ii) provided that the prior for $\mathbf{B}$ encourages sparsity, the resulting classifier is parsimonious and easier to interpret than that of a classifier with dense $\mathbf{B}$.

Figure 1(a) shows a graphical model representation of a PFA module connected to the multi-label classifier in (11), where solid nodes and edges represent PFA module components and dashed lines are specific to the classification model. Combining (6) with (11) allows us to learn the mapping $\mathbf{x}_n \to \mathbf{y}_n$ via the shared first-layer local representation, $\boldsymbol{\theta}_n^{(1)} \circ \mathbf{h}_n^{(1)}$, that encodes topic usage for document n. This sharing mechanism allows the model to learn topics, $\boldsymbol{\Psi}^{(1)}$, and meta-topics, $\{\boldsymbol{\Psi}^{(\ell)}\}_{\ell=2}^{L}$, biased towards discrimination, as opposed to just explaining the data, $\mathbf{x}_n$.

For the deep *multi-modality* model in (7), we learn the mapping $\mathbf{x}_n^{(1)},\ldots,\mathbf{x}_n^{(D)} \to \mathbf{y}_n$ through the first-layer local representations from all modalities, hence $\hat{\boldsymbol{\lambda}}_n = \sum_{i=1}^{D} \mathbf{B}_i(\boldsymbol{\theta}_n^{(i,1)} \circ \mathbf{h}_n^{(i,1)})$, where $\mathbf{B}_i \in \mathbb{R}_+^{C\times K_{(i,1)}}$, for $i = 1,\ldots,D$. In this case, the classifier uses information from all modalities but at the same time biases modality-specific topics, $\boldsymbol{\Psi}^{(i,1)}$, towards discrimination. We call this construction *discriminative* deep multi-modality Poisson factor model. Although other DP-based discriminative topic models have been proposed (Lacoste-Julien et al., 2009; Mcauliffe and Blei, 2008), they rely on approximations in order to combine the topic model, usually LDA, with softmax-based classification approaches.

3.6 Inference

A convenient feature of the model in (6) and (7) is that all its conditional posterior distributions can be written in closed form, due to local conjugacy. In this section, we focus on

Markov chain Monte Carlo (MCMC) via Gibbs sampling for our implementation. In applications where the fully Bayesian treatment becomes prohibitive computationally, Stochastic Variational Inference (SVI) can be used. See Appendix A for details about SVI implementation for models based on PFA modules. Other alternatives for scaling up inference in Bayesian models such as the parameter server (Ho et al., 2013; Li et al., 2014), conditional density filtering (Guhaniyogi et al., 2014) and stochastic gradient-based approaches (Chen et al., 2014; Ding et al., 2014; Welling and Teh, 2011), are also possibile but beyond the scope of this work.

Gibbs sampling for the model in (6) and (7) involves sampling in sequence from the conditional posterior of all the parameters of the model. For instance, for the DPFM in (6), we have $\{\mathbf{\Psi}^{(\ell)}, \boldsymbol{\theta}_n^{(\ell)}, \mathbf{h}_n^{(\ell)}, r_k^{(\ell)}, \boldsymbol{\lambda}^{(0)}\}$, for $\ell = 1, \ldots, L$. For the multi-modality model in (7) we also have to account for modality-specific parameters in (8). The remaining parameters of the model are set to fixed values: $\eta = 1/K$, $b = 0.5$ and $a_0 = b_0 = 1$. We note that priors for η, b, a_0 and b_0 exist that result in Gibbs-style updates, and can be readily incorporated into the model if desired; however, we opted to keep the model as simple as possible, without compromising flexibility. The most unique conditional posteriors for a single PFA module are shown below, without layer index for clarity,

$$\begin{aligned} \boldsymbol{\psi}_k &\sim \text{Dirichlet}(\eta + x_{1k\cdot}, \ldots, \eta + x_{Mk\cdot})\,, \\ \theta_{kn} &\sim \text{Gamma}(r_k h_{kn} + x_{\cdot kn}, b^{-1})\,, \\ h_{kn} &\sim \delta(x_{\cdot kn} = 0)\text{Bernoulli}(\tilde{\pi}_{kn}(\tilde{\pi}_{kn} + 1 - \pi_{kn})^{-1}) + \delta(x_{\cdot kn} = 1)\,, \end{aligned} \tag{12}$$

where

$$x_{mk\cdot} = \sum_{n=1}^{N} x_{mkn}\,, \qquad x_{\cdot kn} = \sum_{m=1}^{M} x_{mkn}\,, \qquad \tilde{\pi}_{kn} = \pi_{kn}(1-b)^{r_k}\,.$$

Complete details, including those for DMPFM and discriminative DMPFM in Sections 3.3 and 3.5, respectively, are provided in Appendix B.

Initialization is done at random from prior distributions, followed by modality-wise and layer-wise fitting (*pre-training*). In the experiments, when pre-training we run 150 Gibbs sampling cycles per layer. We have observed that 50 cycles are usually enough to obtain good initial values of the global parameters of the model, namely $\{\mathbf{\Psi}^{(i,1)}, r_k^{(i,1)}, \mathbf{\Psi}^{(\ell)}, r_k^{(\ell)}, \boldsymbol{\lambda}^{(0)}\}$, for $i = 1, \ldots, D$ and $\ell = 2, \ldots, L$.

3.6.1 Importance of computations scaling with the number of non-zeros

From a practical standpoint, the most important feature of the models in (6) and (7) is that inference does not scale as a function of the size of the total dataset, but as a function of its number of non-zero elements, which is advantageous in cases where the input data is sparse (often the case). For instance, ∼4% of the entries in the dataset described in Section 2 are non-zero. Similar proportions are also observed in datasets traditionally used to bechmark topic models (word documents), such as 20 Newsgroups, Reuters and Wikipedia (details of which are discussed below). Furthermore, this feature also extends to all modalities and layers of the model, regardless of $\{\mathbf{h}_n^{(\ell)}\}$ being latent. Similarly, for the discriminative DMPFM in Section 3.5, inference scales with the number of positive cases in $\{\mathbf{y}_n\}_{n=1}^{N}$, not

CN. This is particularly appealing in cases where C is large and the number of positive cases is small (a patient typically has a small subset of possible illnesses), ∼8% in the dataset described in Section 2.

In order to show that this scaling behavior holds, it is enough to see that by construction, from (3), if $x_{mn} = \sum_{k=1}^{K} x_{mkn} = 0$ (or $z_{mn}^{(\ell)}$ for $\ell > 1$), thus $x_{mkn} = 0$, $\forall k$ with probability 1. Besides, from (5) we see that if $h_{kn} = 0$ then $z_{kn} = 0$ with probability 1. As a result, update equations for all parameters of the model except for $\{\mathbf{h}_n^{(\ell)}\}$, depend only on non-zero elements of $\mathbf{x}_n$ and $\{\mathbf{z}_n^{(\ell)}\}$. Updates for the binary variables can be cheaply obtained in block from $h_{kn}^{(\ell)} \sim \text{Bernoulli}(\pi_{kn}^{(\ell)})$ via $\lambda_{kn}^{(\ell)}$, as previously described.

It is worth mentioning that models based on multinomial or Poisson likelihoods such as LDA (Blei et al., 2003), HDP (Teh et al., 2006), FTM (Williamson et al., 2010) and PFA (Zhou et al., 2012), also enjoy this property (scaling based on number of non-zero observations). However, the recently proposed deep PFA (Gan et al., 2015a) does not use PFA modules on layers other than the first one; it uses SBNs or RBMs that are known to scale with the number of binary variables as opposed to their non-zero elements.

3.7 Related work

3.7.1 Connections to other DP-based topic models

PFA is a nonnegative matrix factorization model with Poisson link, that is closely related to other DP-based models. Specifically, Zhou et al. (2012) showed that by making $p(h_{kn} = 1) = 1$ and letting θ_{kn} have a Dirichlet, instead of a Gamma distribution as in (4), we can recover LDA by using the equivalence between Poisson and multinomial distributions. By looking at (12), we see that PFA and LDA have the same blocked Gibbs updates (Blei et al., 2003), when Dirichlet distributions for θ_{kn} are used. An equivalent analogy for SVI updates (Hoffman et al., 2010) can be derived from the update equations in Appendix A. In Zhou et al. (2012), the authors showed that using the Poisson-gamma representation of the negative binomial distribution and a beta-Bernoulli specification for $p(h_{kn})$ in (4), we can recover the FTM formulation and inference in Williamson et al. (2010). More recently, Zhou and Carin (2015) showed that PFA is comparable to HDP in that the former builds group-specific DPs with normalized gamma processes. A more direct relationship between a three-layer HDP (Teh et al., 2006) and a two-layer version of (6) can be established by grouping count data vectors by categories. In the HDP, three DPs are set for topics, data-dependent topic usage and category-wise topic usage. In our model, $\boldsymbol{\Psi}^{(1)}$ represent K_1 topics, $\boldsymbol{\theta}_n^{(1)} \circ \mathbf{h}_n^{(1)}$ encodes data-vector-wise topic usage and $\boldsymbol{\Psi}^{(2)}$ encodes topic usage for K_2 categories. In HDP, data vectors are assigned to categories *a priori*, but in our model data-vector-category *soft* assignments are estimated and encoded via $\boldsymbol{\theta}_n^{(2)} \circ \mathbf{h}_n^{(2)}$. As a result, the model in (6) is a more flexible alternative to HDP in that it groups data vectors into categories in an unsupervised manner.

3.7.2 Similar models

Non-DP-based deep models for topic modeling employed in the deep learning literature typically utilize RBMs or SBNs as building blocks. For instance, Hinton and Salakhutdinov (2009) and Maaloe et al. (2015) extended RBMs via DBNs to topic modeling. In addition, Srivastava et al. (2013) proposed the over-replicated softmax model, a deep version of RSM that generalizes RBMs.

Recently, Ranganath et al. (2014) proposed a framework for generative deep models using exponential family modules. Although they consider Poisson-Poisson and Gamma-Gamma factorization modules akin to our PFA modules, their model lacks the binary unit linking between layers commonly found in traditional deep models. Further, their inference approach, *black-box* variational inference, is not as conceptually simple and it does not enjoy the scaling with the number of non-zeros of our model.

DPFA, proposed in Gan et al. (2015a), is the model closest to ours. Nevertheless, our proposed model has a number of key differentiating features. (*i*) Both models learn topic correlations by building a multilayer modular representation on top of PFA. Our model uses PFA modules throughout all layers in a conceptually simple and easy to interpret way. DPFA uses Gaussian distributed weight matrices within SBN modules; these are hard to interpret in the context of topic modeling. (*ii*) SBN architectures have the shortcoming of not having block closed-form conditional posteriors for their binary variables, making them difficult to estimate, especially as the number of variables increases. (*iii*) Factor loading matrices in PFA modules have natural shrinkage to counter overfitting, thanks to the Dirichlet prior used for their columns. In SBN-based models, shrinkage has to be added via variable augmentation at the cost of increasing inference complexity. (*iv*) Inference for SBN modules scales with the number of hidden variables in the model, not with the number of non-zero elements, as in our case.

Several deep architectures have been recently proposed for multi-modality problems (Srivastava and Salakhutdinov, 2012, 2014; Sohn et al., 2014). These models use RBMs as building blocks and are traditionally geared towards applications with image (pixel intensities) and text (word counts) modalities. The main goals of these applications are classification based on image and text latent features, and information retrieval, that is, predicting values of one modality given observations of the others. Unlike our discriminative DMPFM and SupDocNADE (Supervised Document Neural Autoregressive Distribution Estimator Zheng et al., 2014) based on SBNs, most existing deep multi-modality models based on RBMs build classifiers as a two-step procedure, not jointly with the generative model as is our case. In its current form, our model does not allow for mixed data-types, however it is not too difficult to extend it to such case, as we can seamlessly use sparse Gaussian factor models (Carvalho et al., 2008; Henao and Winther, 2011) and rank-likelihood factor models (Yuan et al., 2015) as first-layer modules for real and ordinal-valued data, respectively. We leave these extensions as interesting future work.

4. Experiments

In this section we start by presenting benchmark results using well-known corpora for topic models, the goal being to show how DPFM (single modality) compares to related deep models. Next, we present extensive experiments using the motivating data described in Section 2. In particular, we evaluate DPFM and DMPFM in terms of model fit and classification performance. Finally, we analyze the topics estimated by DMPFM. All experiments were conducted on a 2.2GHz desktop machine with 8GB RAM. The code used, implemented in Matlab, will be made publicly available.

4.1 Benchmark corpora

We first evaluate the performance of the basic version of our model, specifically the deep single modality model in (6). For this purpose, we present experiments on three corpora: 20 Newsgroups (20 News), Reuters corpus volume I (RCV1) and Wikipedia (Wiki). 20 News is composed of 18,845 documents and 2,000 words, partitioned into a 11,315 document training set and a 7,531 document test set. RCV1 has 804,414 newswire articles containing 10,000 words. A random 10,000 subset of documents is used for testing. For Wiki, we obtained 10^7 random documents, from which a subset of 1,000 is set aside for testing. Following Hoffman et al. (2010), we keep a vocabulary consisting of 7,702 words taken from the top 10,000 words in the Project Gutenberg Library.

As performance measure, we use held-out perplexity, defined as the geometric mean of the inverse marginal likelihood of every word in the set. We cannot evaluate the intractable marginal for the model in (6), thus we compute the *predictive perplexity* on a 20% subset of the held-out set. The remaining 80% is used to learn document-specific variables of the model, $\{\boldsymbol{\theta}_n^{(\ell)}, \mathbf{h}_n^{(\ell)}\}$, for $n = 1, \ldots, N$ and $\ell = 1, \ldots, L$. The training set is used to estimate the global parameters of the model, $\{\boldsymbol{\Psi}^{(\ell)}, r_k^{(\ell)}, \boldsymbol{\lambda}^{(0)}\}$, for $\ell = 2, \ldots, L$. For PFA-based models, the test perplexity for a single modality can be calculated as (Zhou et al., 2012)

$$\text{perplexity} = \exp\left(-\frac{1}{x_{\cdot\cdot}}\sum_{m=1}^{M}\sum_{n=1}^{N} x_{mn} \log \frac{\sum_{s=1}^{S}\sum_{k=1}^{K}\phi_{mk}^{s}\theta_{kn}^{s}h_{kn}^{s}}{\sum_{s=1}^{S}\sum_{m=1}^{M}\sum_{k=1}^{K}\phi_{mk}^{s}\theta_{kn}^{s}h_{kn}^{s}}\right),$$

where we have omitted modality and layer indices for clarity, S is the total number of collected samples, $x_{\cdot\cdot} = \sum_{m=1}^{M}\sum_{n=1}^{N} x_{mn}$ and x_{mn}, ψ_{mk}, θ_{kn} and h_{kn} are elements of $\mathbf{x}_n$, $\boldsymbol{\Psi}$, $\boldsymbol{\theta}_n$ and $\mathbf{h}_n$, respectively.

We compare our single-modality deep model in (6) (denoted DPFM), against LDA (Blei et al., 2003), FTM (Williamson et al., 2010), RSM (Hinton and Salakhutdinov, 2009), nHDP (Paisley et al., 2015) and DPFA with SBNs (DPFA-SBN) and RBMs (DPFA-RBM) (Gan et al., 2015a). For all these models, we use the settings described in Gan et al. (2015a). Inference methods for RSM and DPFA are contrastive divergence with step size 5 (CD5) and stochastic gradient Nose-Hoover thermostats (SGNHT), respectively. For our model, (after the aforementioned pre-training) we run 3,000 samples, from which the first 2,000 are discarded (burnin). For the Wiki corpus, MCMC-based DPFM is run on a random subset of 10^6 documents.

Table 2 shows results for the corpora being considered. Figures for methods other than DPFM were taken from Gan et al. (2015a). We see that multilayer models (DPFM, DPFA and nHDP) consistently outperform single layer ones (LDA, FTM and RSM), and that DPFM has the best performance across all corpora for models of comparable size. We verified empirically (results not shown) that doubling the number of hidden units, adding a third layer or increasing the number of samples/iterations for DPFM does not significantly change the results in Table 2. As a note on computational complexity, one iteration of the two-layer model on the 20 News corpus takes approximately 2 seconds. For comparison, we also ran the DPFA-SBN model in Gan et al. (2015a) using a two-layer model of the same size; in their case it takes about 24, 4 and 5 seconds to run one iteration using MCMC, conditional density filtering (CDF) and SGNHT, respectively. Runtimes for DPFA-RBM are similar to those of DPFA-SBN.

Model	Method	Size	20 News	RCV1	Wiki
DPFM	MCMC	128-64	**780**	**908**	783
DPFA-SBN	SGNHT	1024-512-256	—	942	**770**
DPFA-SBN	SGNHT	128-64-32	827	1143	876
DPFA-RBM	SGNHT	128-64-32	896	920	942
nHDP	SVI	(10,10,5)	889	1041	932
LDA	Gibbs	128	893	1179	1059
FTM	Gibbs	128	887	1155	991
RSM	CD5	128	877	1171	1001

Table 2: Held-out perplexities for 20 News, RCV1 and Wiki. Size indicates number of topics and/or binary units, accordingly.

4.2 Model fitting

We evaluate the ability of the multi-modality model in (7) to fit the data introduced in Section 2. The data consist of 16,756 patients, and of these 7,892 were used for model fitting and 8,864 for testing. We considered the three modalities discussed above: 1,694 of the entities corresponded to medications (Meds), 1,869 corresponded to laboratory tests (Labs), and 21,013 corresponded to diagnosis and procedure codes (Codes). We filtered out variables with less than 10 occurrences over the entire cohort, reducing the data to 253, 606 and 4,222 entities for Meds, Labs and Codes, respectively. We consider three different models: (*i*) A single-modality approach, in which we treat each modality independently using (6) (denoted Naive1); (*ii*) another single-modality approach, in which all modalities are stacked into one data matrix, then modeled using (6) (denoted Naive2); and (*iii*) the multi-modality approach using (7) (denoted DMPFM). In all cases we collect 1,200 samples after running 1,200 burnin iterations. As the perfomance measure, we report held-out perplexities for each modality on a randomly selected 20% subset (the test set).

Table 3 shows predictive perplexities for different architectures. We consider two- and three-layer specifications (Size) in three different binary unit sizes each, for a total of 6 models. We see that Naive2 and DMPFM consistently outperform Naive1. These results demonstrate that sharing information cross modalities produces a model with a richer correlation structure and improved model fit. We also see that DMPFM performs the best in all configurations, which highlights the importance of modeling correlation structure within and across modalities. In terms of number of layers, we see a modest perplexity improvement going from two to three layers in all cases. This is probably due to the size of the dataset; it is likely that more significant gains will be observed from cohorts with a larger number of variables and patients.

In terms of computational complexity, Naive1 and Naive2 take between 180 and 310 CPU depending on the size of the model; DMPFM takes between 190 and 480 minutes. It is worth mentioning that runtime include model fit, testing and perplexity calculations. In any case, runtimes are deemed reasonable considering the size of the dataset and the complexity of the models being evaluated.

4.3 Multi-label classification

We evaluate the discriminative DMPFM in Section 3.5 on the multi-label classification problem outlined in Section 2. We consider 13 well-known T2DM-related outcomes in Table 1, namely, acute myocardial infarction (AMI), amputation, cardiac catheterization, coronary artery disease, depression, heart failure, chronic kidney disease, neurological disease, obesity, ophthalmic disease, stroke, unstable angina and death. We compare our discriminative DMPFM to discriminative versions of Naive1 and Naive2 based on DPFMs. For a baseline comparison, we use the UKPDS model in (1) and sparse logistic regression (Friedman et al., 2001). For the UKPDS model we estimate model coefficients, $\{b_i\}_{i=0}^{7}$, for each outcome independently in a logistic regression setting. Note that UKPDS was originally intended for coronary heart disease, however we use its covariates (age, sex, race, smoking status, HgbA1c, SPB, TC and HDL) to build classifiers for all outcomes. We also use coronary heart disease interchangeably with coronary artery disease, which is the build up of plaque in the arteries of the heart and results in coronary heart disease. For PFA-based models, we collect 1,200 samples after running 1,200 burnin iterations. As a performance measure, we report area under the receiving operating characteristic (AUCROC) values on the test set (Fawcett, 2006). Provided that all classification tasks are very imbalanced, about 8% positive outcomes in average, we do not report test accuracies. Optimal thresholds can be obtained from ROC curves using outcome-specific prevalence information, if desired. Once threshold values have been selected, accuracies, true positive rates and true negative rates can be readily computed.

Table 4 shows average test AUCs for Naive1, Naive2 and DMPFM and different model architectures. Averages are computed over the 13 outcomes in Table 1. We see from the results for Naive1 that Codes carry significantly more classification power than Med and Lab modalities. Naive2, which combines all modalities into a single data matrix performs consistently better than Naive1, and DMPFM perfroms the best by a considerabe margin, considering the size of the test set. In terms of model size, the largest three-layer model performs the best, closely followed by models of size 96–48–24 and 96–48.

Results in Figure 2a show test AUC values as bars for each outcome individually. We compare DMPFM against two baselines, UKPDS and sparse logistic regression. We see that

	Naive1			Naive2			DMPFM		
Size	Med	Lab	Code	Med	Lab	Code	Med	Lab	Code
64–32	1.930	76.724	210.690	1.930	76.575	208.785	1.865	72.919	194.260
96–48	1.851	76.736	192.851	1.825	76.787	193.782	1.788	72.662	176.737
128–64	1.803	76.538	182.803	1.759	76.495	182.049	1.748	72.415	167.423
64–32–16	1.918	76.648	207.932	1.911	76.400	209.652	1.861	72.773	191.854
96–48–24	1.822	76.967	192.530	1.816	76.660	192.505	1.759	72.531	176.451
128–64–32	1.787	76.556	182.365	1.764	76.528	180.806	**1.730**	**72.364**	**166.759**

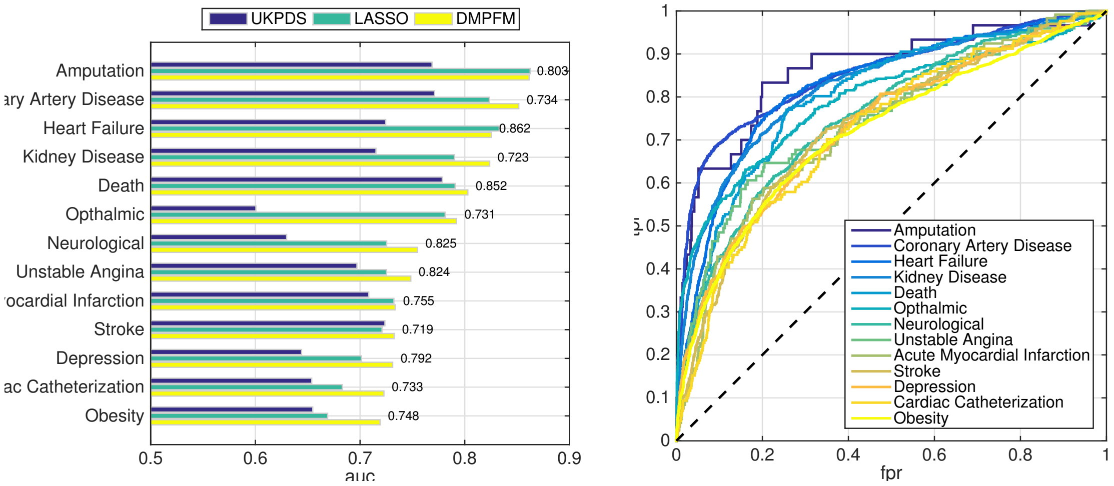
Table 3: Held-out perplexity for EHR data. Size indicates number of topics and/or binary units, accordingly. Naive1 uses one DPFM per modality and Naive2 one DPFM for stacked modalities (Meds, Labs and Codes). Naive2 and DMPFM use all modalities at once but perplexities are computed separately.

		Naive1		Naive2	DMPFM
Size	Med	Lab	Code	All	All
64–32	0.592±0.05	0.594±0.05	0.745±0.06	0.751±0.06	0.771±0.07
96–48	0.596±0.04	0.583±0.05	0.727±0.06	0.750±0.06	0.781±0.06
128–64	0.590±0.04	0.590±0.05	0.725±0.06	0.751±0.06	0.779±0.06
64–32–16	0.601±0.05	0.594±0.05	0.726±0.05	0.742±0.06	0.771±0.06
96–48–24	0.587±0.04	0.588±0.06	0.735±0.06	0.758±0.07	**0.785±0.07**
128–64–32	0.590±0.04	0.588±0.05	0.732±0.05	0.757±0.06	**0.784±0.07**

Table 4: Mean test AUCs with standard deviations over 13 binary classification tasks. Size indicates number of topics and/or binary units, accordingly. Naive1 is one discriminative DPFM per modality and Naive2 is one discriminative DPFM for stacked modalities (Meds, Labs and Codes). Naive2 and DMPFM use all modalities to build classifiers.

DPFM outperforms the others in nearly all classification tasks except for heart failure, in which sparse logistic regression (LASSO) performs best, and amputation, where DMPFM and LASSO perform about the same. In Figure 2a we also show test AUC values obtained by DMPFM sorted in decreasing order, with corresponding ROC curves in Figure 2b.

The outcomes with the greatest predictive power was amputation and the lowest was obesity. Upon further examination these have potentially interesting clinical drivers. For example amputation of limbs in diabetic patients is often the result of longstanding neuropathy and microvascular damage hindering the ability of patients to not only identify injuries



(a)

(b)

Figure 2: Test AUC and ROC curves from multi-label classification experiment. (a) AUC values for UKPDS, sparse logistic regression (LASSO) and DMPFM. Values beside each bar group correspond to AUC values obtained by DMPFM. (b) ROC curves from DMPFM. Each curve represents represents a classification task for an outcome. The dashed line is the AUC of a random classifier.

but also heal (note: the second author is a medical doctor, and provided all medical anaysis). A common clinical scenario involves patients with foot ulcerations that go undetected and result in gangrenous limbs and ultimately amputation. By plotting the classification coefficients for amputation (see Figure 3:top), we identify the top three contributors to this outcome (medications #126, labs #7, and codes #67) in Figure 3:bottom. In medications topic #126, we find the usage of standard diabetes, cholesterol and hypertension medications. Notably we also found Amitriptyline, which can be used to treat patients with diabetic neuropathy. In Labs #7, we identify laboratory tests that would be common in evaluating neuropathy (thiamine deficiency) and a test for a bacterial species common in skin infections, Streptococcus Pyogenes Antigen. This would a be common antigen for a patient with skin infections including foot ulcerations. Lastly, the codes #67 refer to foot ulcers and peripheral vascular disease. Peripheral vascular disease can result from the accumulation of fatty deposits in the vasculature of the extremities and can be exacerbated by the microvascular damage of diabetes. While additional evaluation of topics with high classification coefficients may elicit unexpected predictors of amputation, this initial analysis revealed that the highest scoring topics correlated well with clinical intuition.

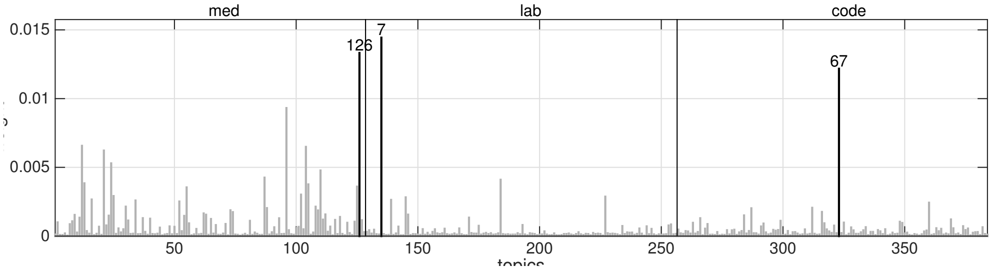
The poor predictive power for obesity likely rests with its prevalence in T2DM populations. Metabolic syndrome, a constellation of symptoms including hyperlipidemia, hypertension and obesity is a strong risk factor for obesity. An examination of contributing first-layer topics reveals medications that would be typical for a patient with symptoms of metabolic syndrome. The main lab first-layer topic shares the same topic #7 as in amputation. Interestingly, morbidly obese patients also share a high incidence of skin infections and pressure induced ulcerations due to their sedentary behaviors. Lastly, the code topics have codes related to abnormal weight gain.

4.4 Analysis of multi-modality model

We also examined the ability of the DPFA model to generate topics that represent intuitive medical concepts. For illustrative purposes, we discuss the intra-modality correlation of first and second level topics (meta-topics), starting with the medications mode and expand to other modalities. We plot the correlations between medication topics in Figure 4. We show first-layer topics (boxes) within a modality. Each box contains the first four words with the highest probability mass in that topic. The topics are connected into meta-topics (blue circles) representing both intra- and inter-modality correlations.

The center of the plot in Figure 4 includes topics (based on lowest layer in the model, touching the data) and meta-topics (based on the second layer in the model) with the highest level of correlation with other topics. Unsurprisingly, the central medicine topic (#3, see Table 5) includes the medication Metformin, a first-line treatment for T2DM, and duloxetine, a treatment for peripheral neuropathy. Interestingly, this topic also includes several steroids including dexamethasone and budesonide, which can induce or worsen T2DM.

To better interpret the correlation structure, we started with meta-topics and explored both intra- and inter-modality correlation. To ease interpretability we focused on meta-topics with fewer connections. We found that many of the meta-topics represented coherent clinical narratives based on the discovered first-layer topics and meta-topics. For example the meta-topic #29 in Figure 5, includes first-layer medicine topics #19 #13 in Table 5.



#7	#126	#67
streptococcus pyogenes ag	magnesium salt	ulcer of other part of foot
hemoglobin.gastrointestinal	aspirin	diabetes with other specifi...
glucose	lisinopril	peripheral vascular disease...
thiamine	metformin	atherosclerosis of native a...
estimated average glucose	acetic acid	rad exam foot comp min 3 vws
ferritin	acetazolamide	interview and evaluation, d...
hemoglobin a1c/hemoglobin.t...	hydromorphone	atherosclerosis of native a...
glucose	calcitriol	ulcer of heel and midfoot
potassium	amitriptyline	gangrene
hemoglobin a2/hemoglobin.total	simvastatin	cellulitis and abscess of f...

Figure 3: Top classification weights and topics associated with amputation. We show the top 10 words (bottom panel) from first-layer topics with the largest 3 classification weights (top panel), namely meds # 126, labs #7 and codes #67.

#3	#13	#19
acid medication	digoxin	clonidine
duloxetine	belladonna alkaloids	simvastatin
metformin	albuterol	valproate
budesonide	duloxetine	colchicine
fluphenazine	acebutolol	fluphenazine
dexamethasone	pseudoephedrine	hydralazine
insulin lispro	azithromycin	omeprazole
atazanavir	cyclosporine	bromfenac
azithromycin	alprostadil	meloxicam
glimepiride	acai berry extract	acid medication

Table 5: Selected topics from medications modality. We show the top 10 words from first-layer topics #3, #13 and #19.

These two collections represent a wide variety of medications used to treat co-morbidities common to T2DM. However, this list also includes opioid pain killers and a chemotherapeutic agent, cyclosporine. While difficult to interpret with only intra-modality correlation, further examination of first-layer topics across modalities that contribute disproportionately to this meta-topic identified laboratory and diagnostics codes which expand the narrative of this meta-topic. A patient weighted heavily with this meta-topic would have laboratory results characterized by hematuria (blood in urine) and prostate specific antigen testing. The combination of these medication and laboratory topics suggests a patient with metastatic

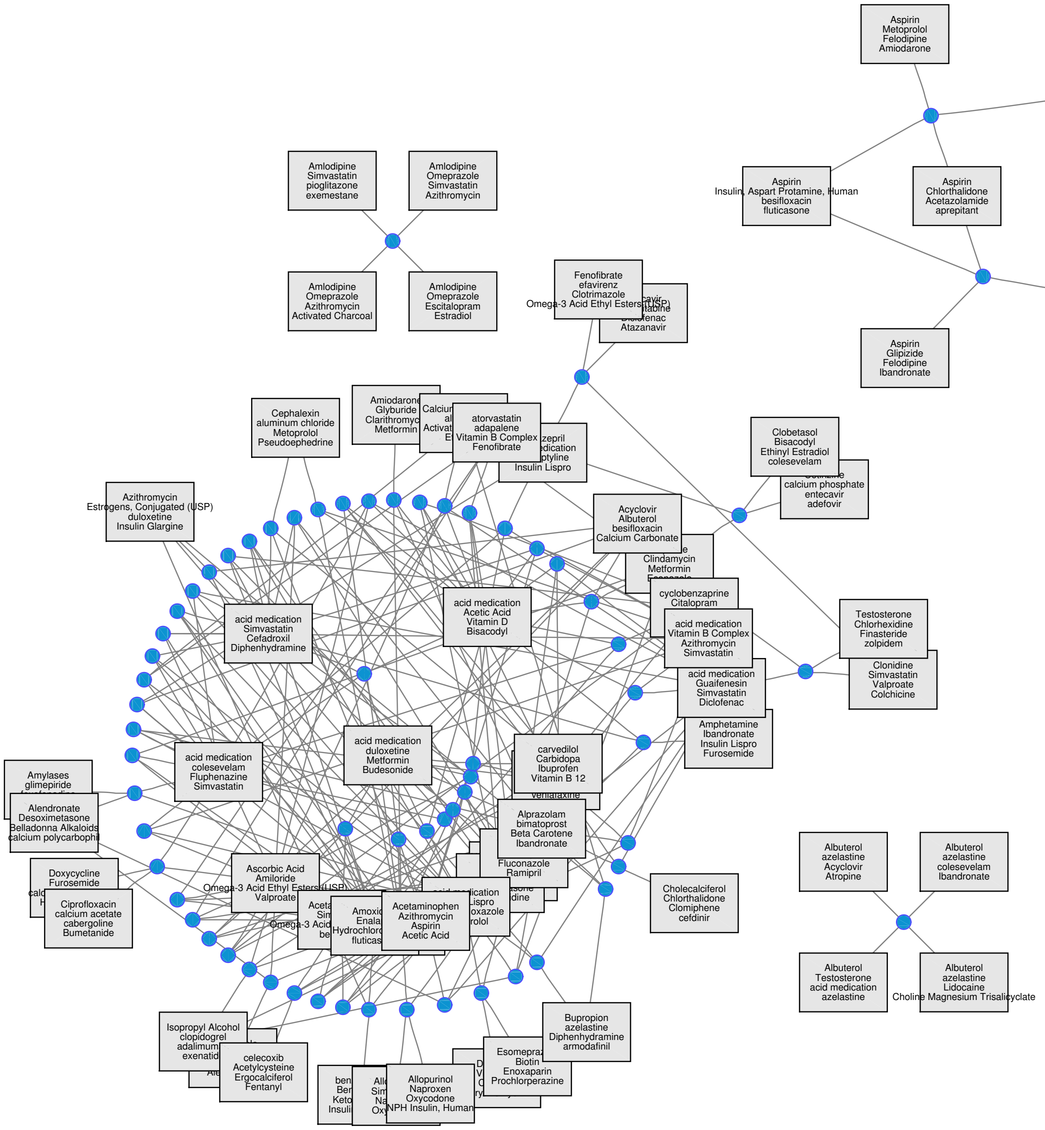


Figure 4: Graph representation obtained for medication interactions. Meta-topic are denoted by circles and first-layer topics as boxes, with word lists corresponding to the top four medications in first-layer topics, $\boldsymbol{\psi}_k^{(1)}$. For clarity, we only show the top four connections between meta-topics and their associated topics.

prostate cancer causing pain and hematuria. This is confirmed when examining the first-layer code topic #84, which includes the code for malignant prostate cancer.

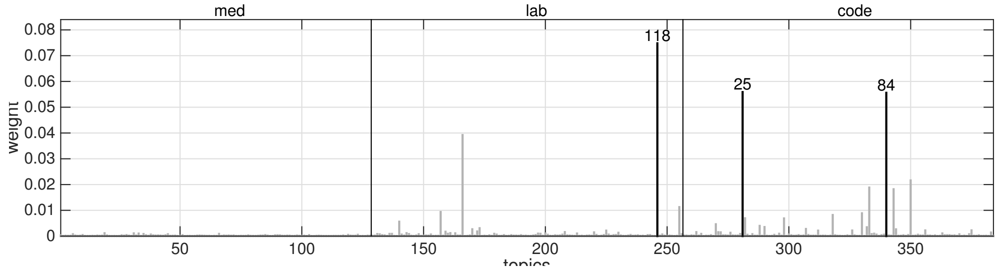
In another example, we explore meta-topic #3 in Figure 6, a topic that does not necessarily make intuitive sense, but could hint at the power of DPFA to identify novel correlations between different data. This meta-topic has two prominent first-layer medication topics. While the first medication topic, #33, contains a mix of hypertensive and antiviral medications, the second topic, #120, includes two notable drugs alprazolam (Xanax) and baclofen, a muscle relaxant. While these two medications may relate to the anxiety, myalgia and insomnia codes, we see in first-layer topic #19 for codes, it would be interesting to explore other first-layer topics contribute to this meta-topic and connect with other conditions identified such as major depressive disorder and chronic pain.

5. Discussion

In 2012, the American Diabetes Association estimated that the economic burden of diabetes in the United States exceeded 245 billion dollars[7]. High-throughput and widely available methods to predict morbidity and mortality outcomes for patients would improve the deployment of medical resources, and may reduce costs through increased preventative care or reduced futile care.

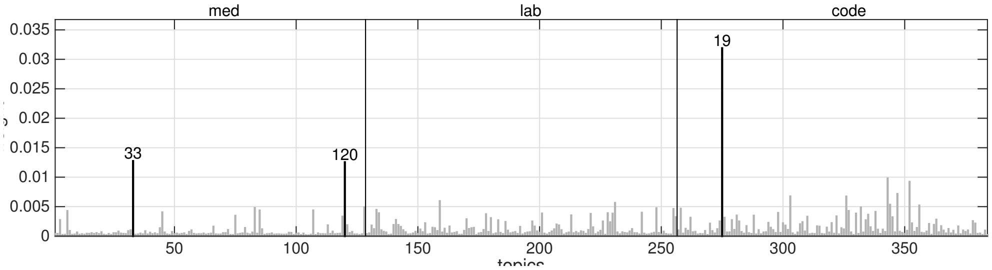
Our initial evaluation of the proposed method illustrated that DMPFM can identify multiple candidate phenotypes from EHR data without expert or user supervision. Further,

[7] See `http://diabetes.org/advocacy/news-events/cost-of-diabetes.html`.



#118	#25	#84
prostate specific ag	routine venipunture for speci	malignant neoplasm of prostate
prostate specific ag.free	routine general medical exa...	impotence of organic origin
prostate specific ag.free/p...	lipid panel	office/outpt visit est pt
testosterone	hemoglobin glycosylated	personal history of maligna...
hemoglobin a1c/hemoglobin.t...	prostate specific antgn totl	elevated prostate specific ...
estimated average glucose	office/outpt visit est pt	interview and evaluation, d...
testosterone.free	reeval healthy est 40-64yrs	other puncture of vein
calcitriol	special screening for malig...	follow-up examination follo...
glomerular filtration rate/...	postop f-up vst incl glob s...	office consult new or est pt
alanine aminotransferase	basic metabolic panel	nocturia

Figure 5: Top weights and topics associated with meta-topic #29. We show the top 10 words (bottom panel) from first-layer topics with the largest 3 meta-topic weights (top panel), namely labs #118 and codes #25 and #84.



#33	#120	#19
calcium carbonate	alprazolam	depressive disorder, not el...
alfuzosin	bimatoprost	office/outpt visit est pt
activated charcoal	beta carotene	anxiety state, unspecified
etodolac	ibandronate	postop f-up vst incl glob s...
donepezil	amiodarone	myalgia and myositis, unspe...
adefovir	azelastine	dysthymic disorder
benazepril	clotrimazole	psych insight off/op 45-50 m
aspirin	cephalexin	insomnia, unspecified
vitamin b 12	irbesartan	chronic pain syndrome
amantadine	baclofen	major depressive disorder, ...

Figure 6: Top weights and topics associated with meta-topic #3. We show the top 10 words (bottom panel) from first-layer topics with the largest 3 meta-topic weights (top panel), namely labs #19 and codes #33 and #120.

these candidate topics, when defined in the context of a classification task, can significantly outperform current benchmarks for risk prediction derived from large-scale clinical studies. This was perhaps unsurprising as we are able to utilize much richer datasets (in both number of clinical variables and patient numbers) than most clinical studies. We are also able to estimate the important factors directly from the data, thus minimizing bias on the behalf of the original study designers.

Yet despite these encouraging findings, many challenges remain before such high-throughput phenotyping efforts can be used in a clinical setting. First, the clinical evaluation of DMPFM for EHR relied on a single clinical expert (second author) to perform the data reconciliation and evaluate the topic groupings. This evaluation is subject to bias, and a more extensive study involving a panel of clinicians is necessary to validate this method's robustness. Second, as discussed above, some of the clinical phenotypes are not easily interpreted. Although our method encourages sparsity there remains a high level of correlation both within and between topics. These topics have many interconnections that remain to be fully explored. In addition, the appropriate metrics to evaluate words (entities), topics and meta-topics for clinical applications requires further research. Third, even in cases where we can generate a narrative around candidate phenotypes, many topics still contained words/entities from each modality that appeared irrelevant to the larger meta-topic. Additional research is needed to establish the clinical relevance of these words/entities. Fourth, although we present a generative process for DMPFM, we did not perform experiments exploring the generation of topics weights for new patients. It would also be interesting to explore the number of patients that would be required to generate a fully comprehensive but sparse set of topics for any given patient population. Lastly, although it appears that we can generate meta-topics that represent patients, we did not perform case review of these patients that review those topics/meta-topics. It would be necessary to perform off-line review by physicians to establish clinical correlation between computational phenotypes and the true patient status.

The DMPFM is an extensible model applicable to any data modalities that can be represented with count data. This would naturally extend to free-text physician and nursing notes (in this case the counts are of actual words) as well as notes from specialty services, such as radiology and pathology. With the panoply of additional data that is contained with the medical record, we are confident that we can develop improved representations of patient traits that may lead to better diagnosis and treatment outcomes.

Acknowledgments

We would like to acknowledge support for this project from the Information Initiative at Duke, Duke Clinical Research Institute and Duke Medicine.

Appendix A. Stochastic variational inference

SVI is a scalable algorithm for approximating posterior distributions consisting of EM-style local-global updates, in which subsets of a dataset (*mini-batches*) are used to update in closed-form the variational parameters controlling both the local and global structure of the model in an iterative fashion Hoffman et al. (2013). This is done by using stochastic optimization with noisy natural gradients to optimize the variational objective function. Additional details and theoretical foundations of SVI can be found in Hoffman et al. (2013).

In practice the algorithm proceeds as follows, where again we have omitted the layer index for clarity: (*i*) let $\{\mathbf{\Psi}^{(t)}, r_k^{(t)}, \boldsymbol{\lambda}^{(t)}\}$ be the global variables at iteration t. (*ii*) Sample a mini-batch from the full dataset. (*iii*) Compute updates for the variational parameters of the local variables using

$$\begin{aligned}
\phi_{mkn} &\propto \exp(\mathbb{E}[\log\psi_{mk}] + \mathbb{E}[\log\theta_{kn}])\,, \\
\theta_{kn} &\sim \text{Gamma}(\mathbb{E}[r_k]\mathbb{E}[h_{kn}] + \sum_{m=1}^{M} x_{mn}\phi_{mkn}, b^{-1})\,, \\
h_{kn} &\sim \mathbb{E}[p(x_{\cdot kn}=0)]\text{Bernoulli}(\mathbb{E}[\tilde{\pi}_{kn}](\mathbb{E}[\tilde{\pi}_{kn}] + 1 - \mathbb{E}[\pi_{kn}])]^{-1}) + \mathbb{E}[p(x_{\cdot kn}=1)]
\end{aligned}$$

where $\mathbb{E}[x_{mkn}] = \phi_{mkn}$ and $\mathbb{E}[\tilde{\pi}_{kn}] = \mathbb{E}[\pi_{kn}](1-b_n)^{\mathbb{E}[r_k]}$. In practice, expectations for θ_{kn} and h_{kn} are computed in log-domain. (*iv*) Compute a local update for the variational parameters of the global variables (only $\mathbf{\Psi}$ is shown) using

$$\widehat{\psi}_{mk} = \eta + \frac{N}{N_B}\sum_{n=1}^{N_B} x_{mn}\phi_{mkn}\,, \tag{13}$$

where N and N_B are sizes of the corpus and mini-batch, respectively. Finally, we update the global variables as $\boldsymbol{\psi}_k^{(t+1)} = (1-\rho_t)\boldsymbol{\psi}_k^{(t)} + \rho_t\widehat{\boldsymbol{\psi}}_k$, where $\rho_t = (t+\tau)^{-\kappa}$. The forgetting rate,

$\kappa \in (0.5, 1]$ controls how fast previous information is forgotten and the delay, $\tau \geq 0$, down-weights early iterations. These conditions for κ and τ guarantee that the iterative algorithm converges to a local optimum of the variational objective function. In the experiments, we set $\kappa = 0.7$ and $\tau = 128$. Additional details of the SVI algorithm for the model in (6) are given in Appendix B.

Appendix B. Inference details

MCMC

Conditional posteriors (layer index omitted for clarity):

$$\begin{aligned}
\boldsymbol{\psi}_k &\sim \text{Dirichlet}(\eta + x_{1k\cdot}, \ldots, \eta + x_{Mk\cdot})\,, \\
\theta_{kn} &\sim \text{Gamma}(r_k h_{kn} + x_{\cdot kn}, b^{-1})\,, \\
h_{kn} &\sim \delta(x_{\cdot kn} = 0)\text{Bernoulli}(\tilde{\pi}_{kn}(\tilde{\pi}_{kn} + 1 - \pi_{kn})^{-1}) + \delta(x_{\cdot kn} = 1)\,, \\
r_k &\sim \text{Gamma}\left(1 + \sum_n u_{kn}, 1 - \sum_n h_{kn} \log(1 - b)\right)\,, \\
z_{kn} &\sim \delta(h_{kn} = 1)\text{Poisson}_+(\tilde{\lambda}_{kn})\,,
\end{aligned}$$

where $\text{Poisson}_+(\cdot)$ is the zero-truncated Poisson distribution and

$$\begin{aligned}
x_{mk\cdot} &= \sum_{n=1}^{N} x_{mkn}\,, \\
x_{\cdot kn} &= \sum_{m=1}^{M} x_{mkn}\,, \\
\tilde{\pi}_{kn} &= \pi_{kn}(1 - b)^{r_k}\,, \\
u_{kn} &= \sum_{j=1}^{x_{\cdot kn}} u_{knj}\,, \qquad u_{knj} \sim \text{Bernoulli}\left(\frac{r_k}{r_k + j - 1}\right)\,.
\end{aligned} \tag{14}$$

Note that for multilayer models, $\pi_{kn}^{(\ell)} = 1 - \exp(\lambda_{kn}^{(\ell+1)})$. The data augmentation scheme for r_k via u_{kn} is described in Zhou and Carin (2015).

For the discriminative DPFM, lets denote latent counts for $\widehat{y}_n$ as $\widehat{x}_{ckn}$, with summaries analogous to (14), as $\widehat{x}_{ck\cdot}$ and $\widehat{x}_{\cdot kn}$. Then,

$$\begin{aligned}
\mathbf{b}_k &\sim \text{Dirichlet}(\zeta + \widehat{x}_{1k\cdot}, \ldots, \zeta + \widehat{x}_{Ck\cdot})\,, \\
\theta_{kn} &\sim \text{Gamma}(r_k h_{kn} + x_{\cdot kn} + \widehat{x}_{\cdot kn}, b^{-1})\,, \\
h_{kn} &\sim \delta(x_{\cdot kn} = 0 \wedge \widehat{x}_{\cdot kn} = 0)\text{Bernoulli}(\tilde{\pi}_{kn}(\tilde{\pi}_{kn} + 1 - \pi_{kn})^{-1}) + \delta(x_{\cdot kn} = 1 \vee \widehat{x}_{\cdot kn} = 1)\,.
\end{aligned}$$

Provided that $\boldsymbol{\theta}_n$ and $\mathbf{h}_n$ are shared by two PFA modules, one for the count data, $\mathbf{x}_n$, and the other for the labels, $\widehat{y}_n$, their conditional posteriors are functions of latent counts coming from both sources, $x_{\cdot kn}$ and $\widehat{x}_{\cdot kn}$, respectively.

SVI

Variational parameter updates using (layer index omitted for clarity):

$$
\begin{aligned}
\phi_{mkn} &\propto \exp(\mathbb{E}[\log \psi_{mk}] + \mathbb{E}[\log \theta_{kn}])\,, \\
\theta_{kn} &\sim \mathrm{Gamma}(\mathbb{E}[r_k]\mathbb{E}[h_{kn}] + \sum_{m=1}^{M} x_{mn}\phi_{mkn}, b^{-1})\,, \\
h_{kn} &\sim \mathbb{E}[p(x_{\cdot kn} = 0)]\mathrm{Bernoulli}(\mathbb{E}[\tilde{\pi}_{kn}](\mathbb{E}[\tilde{\pi}_{kn}] + 1 - \mathbb{E}[\pi_{kn})]^{-1}) + \mathbb{E}[p(x_{\cdot kn} = 1)]\,, \\
r_k &\sim \mathrm{Gamma}\left(1 + \sum_n \mathbb{E}[u_{kn}], 1 - \sum_n \mathbb{E}[p(h_{kn} = 1)] \log(1 - b)\right)\,, \\
z_{kn} &\sim \mathbb{E}[p(h_{kn} = 1)]\mathrm{Poisson}_{+}(\tilde{\lambda}_{kn})\,,
\end{aligned}
$$

where

$$
\mathbb{E}[x_{mkn}] = \phi_{mkn}\,, \quad \mathbb{E}[\tilde{\pi}_{kn}] = \mathbb{E}[\pi_{kn}](1 - b_n)^{\mathbb{E}[r_k]}\,, \quad \mathbb{E}[u_{kn}] = \sum_{j=1}^{x_{\cdot kn}} \mathbb{E}[r_k](\mathbb{E}[r_k] + j - 1)^{-1}\,.
$$

References

American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care*, 37(1):81–90, 2014.

David M. Blei and John D. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1(1):17–35, 2007.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(993–1022), 2003.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems*, 2004.

Carlos M. Carvalho, Jeffrey Chang, Joseph E. Lucas, Joseph R. Nevins, Quanli Wang, and Mike West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

Tianqi Chen, Emily B. Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, 2014.

Yukun Chen, Robert J. Carroll, Eugenia R. McPeek Hinz, Anushi Shah, Anne E. Eyler, Joshua C. Denny, and Hua Xu. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association*, 20(e2):e253–9, 2013.

P. M. Clarke, A. M. Gray, A. Briggs, A. J. Farmer, P. Fenn, R. J. Stevenson, D. R. Matthews, I. M. Stratton, and R. R. Holman. A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) outcomes model (UKPDS no. 68). *Diabetologia*, 47(10):1747–1759, 2004.

David Collett. *Modelling binary data.* CRC Press, 2002.

Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D. Skeel, and Hartmut Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in Neural Information Processing Systems*, 2014.

Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning.* Springer series in statistics Springer, Berlin, 2001.

Zhe Gan, Changyou Chen, Ricardo Henao, David Carlson, and Lawrence Carin. Scalable deep Poisson factor analysis for topic modeling. In *International Conference on Machine Learning*, 2015a.

Zhe Gan, Ricardo Henao, David Carlson, and Lawrence Carin. Learning deep sigmoid belief networks with data augmentation. In *International Conference on Artificial Intelligence and Statistics*, 2015b.

Rajarshi Guhaniyogi, Shaan Qamar, and David B. Dunson. Bayesian conditional density filtering. *arXiv:1401.3632*, 2014.

Ricardo Henao and Ole Winther. Sparse linear identifiable multivariate modeling. *Journal of Machine Learning Research*, 12:863–905, 2011.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–800, 2002.

Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Replicated softmax: an undirected topic model. In *Advances in Neural Information Processing Systems*, 2009.

Joyce C. Ho, Joydeep Ghosh, Steve R. Steinhubl, Walter F. Stewart, Joshua C. Denny, Bradley A. Malin, and Jimeng Sun. Limestone: high-throughput candidate phenotype generation via tensor factorization. *Journal of biomedical informatics*, 52:199–211, 2014a.

Joyce C. Ho, Joydeep Ghosh, and Jimeng Sun. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *International Conference on Knowledge Discovery and Data Mining*, 2014b.

Qirong Ho, James Cipar, Henggang Cui, Seunghak Lee, Jin Kyu Kim, Phillip B. Gibbons, Garth A. Gibson, Greg Ganger, and Eric P. Xing. More effective distributed ML via a stale synchronous parallel parameter server. In *Advances in Neural Information Processing Systems*, 2013.

Matthew Hoffman, Francis R. Bach, and David M. Blei. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, 2010.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

George Hripcsak and David J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–21, 2013.

P. King, I. Peacock, and R. Donnelly. The UK prospective diabetes study (UKPDS): clinical and therapeutic implications for type 2 diabetes. *British journal of clinical pharmacology*, 48(5):643–8, 1999.

Simon Lacoste-Julien, Fei Sha, and Michael I Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems*, 2009.

Hugo Larochelle and Stanislas Lauly. A neural autoregressive topic model. In *Advances in Neural Information Processing Systems*, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Mu Li, David G. Andersen, Alex J. Smola, and Kai Yu. Communication efficient distributed machine learning with the parameter server. In *Advances in Neural Information Processing Systems*, 2014.

Shou-En Lu, Gloria L. Beckles, Jesse C. Crosson, Dorian Bilik, Andrew J. Karter, Robert B. Gerzoff, Yong Lin, Sonja V. Ross, Laura N. McEwen, Beth E. Waitzfelder, David Marrero, Norman Lasser, and Arleen F. Brown. Evaluation of risk equations for prediction of short-term coronary heart disease events in patients with long-standing type 2 diabetes: the translating research into action for diabetes (TRIAD) study. *BMC Endocrine Disorders*, 12(12):1–10, 2012.

Lars Maaloe, Morten Arngren, and Ole Winther. Deep belief nets for topic modeling. *arXiv:1501.04325*, 2015.

Ravi K. Mareedu, Falgun M. Modhia, Elenita I. Kanin, James G. Linneman, Terrie Kitchner, Catherine A. McCarty, Ronald M. Krauss, and Russell A. Wilke. Use of an electronic medical record to characterize cases of intermediate statin-induced muscle toxicity. *Preventive cardiology*, 12(2):88–94, 2009.

Jon D. Mcauliffe and David M. Blei. Supervised topic models. In *Advances in Neural Information Processing Systems*, 2008.

Patricia A. Metcalf, Susan Wells, Robert K. R. Scragg, and Rod Jackson. Comparison of three different methods of assessing cardiovascular disease risk in new zealanders with type 2 diabetes mellitus. *The New Zealand medical journal*, 121(1281):49–57, 2008.

Radford M. Neal. Connectionist learning of belief networks. *Artificial Intelligence*, 56(1): 71–113, 1992.

Stuart J. Nelson, Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore. Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association*, 18(4):441–8, 2011.

Katherine M. Newton, Peggy L. Peissig, Abel Ngo Kho, Suzette J. Bielinski, Richard L. Berg, Vidhu Choudhary, Melissa Basford, Christopher G Chute, Iftikhar J. Kullo, Rongling Li, Jennifer A. Pacheco, Luke V. Rasmussen, Leslie Spangler, and Joshua C. Denny. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association*, 20(e1):e147–54, 2013.

John Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. Nested hierarchical Dirichlet processes. *Pattern Analysis and Machine Intelligence*, 37(2):256–270, 2015.

Walter W. Piegorsch. Complementary log regression for generalized linear models. *The American Statistician*, 46(2):94–99, 1992.

Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David M. Blei. Deep exponential families. In *International Conference on Artificial Intelligence and Statistics*, 2014.

Rachel L. Richesson, Shelley A. Rusincovitch, Douglas Wixted, Bryan C. Batch, Mark N. Feinglos, Marie Lynn Miranda, W. Ed Hammond, Robert M. Califf, and Susan E. Spratt. A comparison of phenotype definitions for diabetes mellitus. *Journal of the American Medical Informatics Association*, 20(e2):e319–26, 2013.

Rebecca K. Simmons, Ruth L. Coleman, Hermione C. Price, Rury R. Holman, Kay T. Khaw, Nicholas J. Wareham, and Simon J. Griffin. Performance of the UK prospective diabetes study risk engine and the framingham risk equations in estimating cardiovascular disease in the EPIC-norfolk cohort. *Diabetes Care*, 32(4):708–13, 2009.

Kihyuk Sohn, Wenling Shang, and Honglak Lee. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pages 2141–2149, 2014.

Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems*, 2012.

Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15:2949–2980, 2014.

Nitish Srivastava, Ruslan Salakhutdinov, and Geoffrey E. Hinton. Modeling documents with deep Boltzmann machines. In *Uncertainty in Artificial Intelligence*, 2013.

R. J. Stevens, V. Kothari, A. I. Adler, I. M. Stratton, and United Kingdom Prospective Diabetes Study (UKPDS) Group. The UKPDS risk engine: a model for the risk of coronary heart disease in type ii diabetes (UKPDS 56). *Clinical science (London)*, 101 (6):671–9, 2001.

Libo Tao, Edward C. F. Wilson, Simon J. Griffin, Rebecca K. Simmons, and ADDITION-Europe study team. Performance of the UKPDS outcomes model for prediction of myocardial infarction and stroke in the ADDITION-Europe trial cohort. *Value Health*, 16 (6):1074–80, 2013.

Yee W. Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

S. van Dieren, L. M. Peelen, U. Nöthlings, Y. T. van der Schouw, G. E. Rutten, A. M. Spijkerman, D. L. van der A, D. Sluik, H. Boeing, K. G. Moons, and J. W. Beulens. External validation of the UK prospective diabetes study (UKPDS) risk engine in patients with type 2 diabetes. *Diabetologia*, 54(2):264–70, 2011.

Daniel J. Vreeman, John Hook, and Brian E. Dixon. Learning from the crowd while mapping to LOINC. *Journal of the American Medical Informatics Association*, 2015.

Max Welling and Yee W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, 2011.

Sinead Williamson, Chong Wang, Katherine Heller, and David Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In *International Conference on Machine Learning*, 2010.

P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97 (18):1837–47, May 1998.

Xin Yuan, Ricardo Henao, Ephraim L. Tsalik, Raymond Langley, and Lawrence Carin. Non-gaussian discriminative factor models via the max-margin rank-likelihood. In *International Conference on Machine Learning*, 2015.

Yin Zheng, Yu J. Zhang, and Hugo Larochelle. Topic modeling of multimodal data: an autoregressive approach. In *Computer Vision and Pattern Recognition*, 2014.

Mingyuan Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *International Conference on Artificial Intelligence and Statistics*, 2015.

Mingyuan Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *Pattern Analysis and Machine Intelligence*, 37(2):307–320, 2015.

Mingyuan Zhou, Lauren Hannah, David Dunson, and Lawrence Carin. Beta-negative binomial process and Poisson factor analysis. In *International Conference on Artificial Intelligence and Statistics*, 2012.