Supervised Topic Modeling in Clinical Text

Perotte A **Li Y** **Pivovarov R** **Weiskopf N** **Wood F**
Columbia University, New York, NY 10027, USA
{{ajp9009, yil7003, rip7002, ngw7001}@dbmi., fwood@stat.}columbia.edu

Abstract

Current medical record keeping technology relies heavily upon human capacity to effectively summarize and infer information from free-text physician notes. We propose a novel method to suggest diagnostic code assignment for patient visits, based upon narrative medical notes. We applied a supervised latent Dirichlet allocation model to a corpora of free-text medical notes from NewYork - Presbyterian Hospital to infer a set of specific ICD-9 codes for each patient note. Evaluation of the predictions were conducted by comparison to a gold-standard set of ICD-9s assigned to a set of patient notes.

1 Introduction

Despite the growing emphasis on meaningful use of technology in medicine, many aspects of medical record-keeping remain a manual process. Diagnostic coding for billing and insurance purposes is often handled by professional medical coders who must explore a patient's extensive clinical record before assigning the proper codes. So while electronic health records (EHRs) should be adopted by most medical institutions within the next several years, largely due to the provisions of HITECH under the American Recovery and Reinvestment Act [4], there has been little movement forward in automating medical coding.

In this paper we describe the use of a topic model based on supervised latent Dirichlet allocation (sLDA) to identify topics within narrative discharge summaries and to automate the assignment of diagnostic codes, specifically International Classification of Disease 9th Revision, Clinical Modification (ICD-9-CM) codes. There are a number of advantages to this approach. First, manually coding diagnoses is a time-consuming and notoriously unreliable process. Many diagnoses are omitted in the final record, and a high error rate is found even in the principal diagnoses [15].

While there have been some attempts to automatically classify groups of patients as a potential preliminary step to ICD-9 code assignment [14, 8, 13, 5], fully automatic assignment of ICD-9 codes to medical text became a more prevalent research topic only in the last few years. A subset of earlier work proposed various methods on small corpora, based on a few specific diseases [12] but the most recent and promising work on the subject was inspired by the 2007 Medical NLP Challenge: International Challenge: Classifying Clinical Free Text Using Natural Language Processing (website). Most of the classification strategies included word matching and rule-based algorithms. [9, 6, 7]. The data set given to the participants consisted only of documents that were 1-2 lines each and all of the documents were radiology reports - clearly limiting the scope of potential ICD-9 codes which could be assigned. The only paper which has attempted to work with a document scope as large as ours was the 2008 Lita et al publication [11]. Lita proposed support vector machine and bayesian ridge regression methods to assign appropriate labels to the documents but did not utilize the ICD-9 hierarchy to leverage more comprehensive predictions.

An automated process would ideally produce a more complete and accurate diagnosis lists. Also, this model will reveal information about the medical records themselves. For example, we may gain an understanding of what a specific code actually means in terms of clinical narratives. Similarly,

viewing the distribution of topics over discharge summaries may reveal information about the latent structure of clinician documentation. Lastly, the sLDA model would provide a novel approach to dealing with the problem of high dimensionality when representing narrative text in a vector space specifically by reducing dimensions from an entire vocabulary of potentially tens of thousands of words to a set of several dozen topics.

2 Methods

2.1 Data

Our data set was gathered from the clinical data warehouse of NewYork - Presbyterian Hospital. The data consisted of free-text discharge summaries and their respective ICD-9 codes. A discharge summary is a clinical report prepared by a physician or other health professional at the conclusion of a hospital stay or series of treatments. The note outlines the patients chief complaint, diagnostic findings, therapy administered, patients response to the chosen therapy, the treatment plan and the recommendations upon discharge. The ICD-9 codes used to structure the discharge summary data are part of a controlled terminology which is the international standard diagnostic classification for epidemiological, health management, and clinical purposes (http://www.who.int/classifications/icd/en/). The codes are classified in a rooted-tree structure, with each edge representing an is-a relationship between parent and child, such that the parent diagnosis subsumes the child diagnosis. In the hospital, ICD-9 codes are generated manually by trained medical coders, who review all the information in the discharge summary. For the purposes of sLDA, ICD-9 codes will be used as labels for discharge summaries.

We worked within the guidelines of the Health Insurance Portability and Accountability Act (HIPAA), which protects patient privacy and the security of potentially identifying medical material, known as personal health information (PHI). HIPAA covers any information within a medical record that was created, used, or disclosed during the course of providing a health care service and that can be used to identify an individual. Before beginning data processing, we generated a PHI-free dataset (see Data Pre-processing below).

2.2 Pre-Processing

Patient discharge summaries and their associated ICD-9 diagnoses are stored in two different places in the NewYork - Presbyterian data warehouse, and so had to be linked together before being fed into the sLDA algorithm. Each discharge note and set of diagnoses were assigned a patient unique identifier (PUID) and a visit unique identifier (VUID), allowing the two types of data to be linked.

Natural Language Processsing (NLP) techniques were used to process the free-text discharge summaries. First, the Natural Language Toolkit (http://www.nltk.org/) was used to tokenize the text. Next, feature selection was performed using a term frequency - inverse document frequency (TF-IDF) algorithm on the entire document set and sorting the words by their TF-IDF values. The top 10,000 words were manually evaluated to eliminate all potentially identifying information. Finally, each discharge summary was converted to a bag-of-words, listing the frequencies of the remaining, free of protected health information, top 10,000 words.

Preparation of the diagnostic codes involved inference over the ICD-9 hierarchy. The is-a relationships of the hierarchy allowed us to make two important assumptions. First, if a diagnosis was observed to be present, all of its ancestors could be assumed to be present as well (e.g., if a patient had malignant hypertension, it could be assumed that they also had essential hypertension. Second, if a diagnosis was observed to be absent, it could be assumed that all of its descendants were also absent (e.g. if a patient did not have essential hypertension, it could be assumed that they did not have malignant hypertension). Unfortunately, ICD-9 code observations never include observations of disease absence. ICD-9 codes are only documented when the condition is observed to be present. Additionally, ICD-9 codes are known to have relatively low sensitivity; conditions that are present are often not documented in a set of ICD-9 codes (Surjan, 1999). Given these facts, we made the following assumptions regarding each visit: recorded diagnoses and their ancestors were labeled as true; diagnoses that were observed at some time for a patient but not at the current visit were labeled as unobserved; and diagnoses that had never been listed for a patient were labeled as false for all of that patients visits. This last assumption captures the belief that parts of the ICD-9 hierarchy that are

never observed for a particular patient are almost certain to be false. Additionally, for computational purposes, we decided not to include an ICD-9 code at all if neither it nor one of its descendants had been assigned to a patient in any of the records in our dataset.

2.3 Supervised Topic Models

Latent dirichlet allocation (LDA) is a generative probabilistic model of corpora that represents documents as a mixed membership bag-of-words. Also known as topic models, these models infer the latent structure, or topics, of documents in a corpus. Each document is represented as a collection of words, generated from a set of topic assignments (one for each word), where each topic assignment is drawn from a distribution over topics [3].

This model, supervised latent dirichlet allocation (sLDA), builds on LDA by incorporating an exponential family response variable. Although there are many models for making predictions based on free text, sLDA is unique in that it is a generative model, it represents documents as a mixed-membership, and constrains the inference of the latent structure of the documents by its predictability of the response variable. In other words, sLDA infers topics such that the model is capable of a high predictive likelihood for words in a document and the response variable associated with a document. This approach has been shown to outperform both LASSO (L1 regularized least squares regression) and LDA followed by least squares regression [2].

Here, we augment the sLDA model such that the supervised signal is distribution over the ICD-9 code tree, which is an is-a hierarchy [1]. An is-a hierarchy is represented by the tree data structure where each node has only a single parent and nodes cannot be parents of ancestors (ie. there are no loops). In this particular case, the ICD-9 code hierarchy is also partially a prefix trie where the labels for certain nodes are prefixes for child nodes. Given that this rule does not apply to all nodes in the hierarchy, we did not use this feature to determine the structure of the hierarchy. Instead we acquired a dataset that explicitly defined the relationships between the nodes of the hierarchy [1]. In documentation of ICD-9 codes for billing purposes, only a subset of the nodes can be used, however the nodes higher in the hierarchy contain semantic information about the categories of codes that are their descendants. For this reason, we included these nodes in our model.

2.4 Generative Model

Given the number of topics, K, the global prior over topic proportions, α', and the prior over topics, γ, the generative process for documents and responses is as follows:

1. For each topic:
   (a) Draw a distribution over words $\beta_k \sim Dir\,(u, \gamma)$
2. For each ICD9 Code:
   (a) Draw regression coefficient $\eta_i \mid \mu, \sigma \sim \mathcal{N}\,(\mu, \sigma)$
3. Draw a prior over topic proportions $m \mid \alpha' \sim Dir\,(u, \alpha')$
4. For each document:
   (a) Draw topic proportions $\theta_d \mid \alpha \sim Dir\,(m, \alpha)$
   (b) For each word:
      i. Draw topic assignment $z_{n,d} \mid \theta_d \sim Mult\,(\theta_d)$
      ii. Draw word $w_{n,d} \mid z_{n,d}, \beta_{1:K} \sim Mult\,(\beta_{z_n})$
   (c) For each ICD-9 code from the root of the hierarchy and recursively descending the tree:
      i. Draw a response variable $y_i \mid \bar{z}, \eta_i, y_{parent} \sim \Phi(\eta_i^T \bar{z}, y_{parent})$ where $\bar{z} = N^{-1} \sum_{n=1}^{N} z_n$ and Φ refers to a conditional probit model.

We will employ a data augmentation scheme with auxiliary variables a_i in the probit model where:

$$y_i \sim \begin{cases} 1, & a_i > 0 \; and \; y_{parent} = 1 \\ 0, & otherwise \end{cases} \tag{1}$$

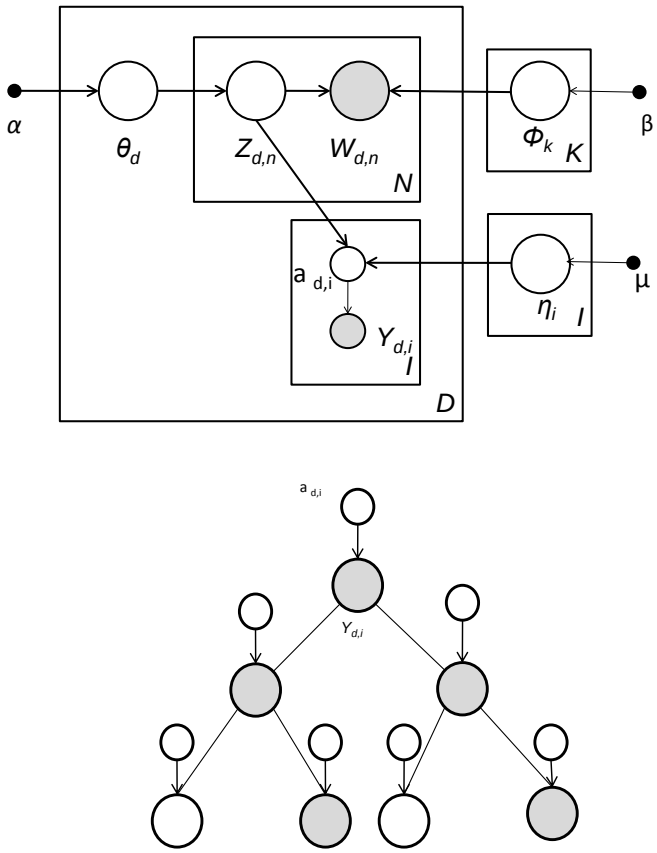

Figure 1: adapted sLDA model

$$a_i \sim \mathcal{N}\left(\eta_i^T \bar{z}, 1\right) \tag{2}$$

2.5 Posterior Inference

Given an observation of a set of ICD-9 codes and a document, the posterior distribution for the latent variables is given by Equation 4.

The denominator for this distribution is the marginal probability of the data and cannot be solved in closed form. This is often the case in evaluating posterior distributions of non-trivial probabilistic models. We will appeal to one of the common methods for approximating posterior distributions in the face of intractable normalization factors: Markov chain Monte Carlo (MCMC) sampling. Since in this model it is possible to sample from the conditional distributions for all variables we will use the Gibbs sampling algorithm to obtain an approximation to this posterior. In particular, we will derive a collapsed Gibbs sampler for the supervised topic model.

2.6 Rao-Blackwellization

To derive the Gibbs sampler in general, we integrate over the parameters θ and $\phi_{1:K}$ resulting in the collapsed joint distribution shown in Equations 5-8.

$$p\left(\theta, z_{1:N} \mid w_{1:N}, y_{1:I}, \phi_{1:K}, \eta_{1:I}, \alpha, \beta, \mu, \xi\right) = \frac{p\left(\theta \mid \alpha\right)\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right)}{\int_\theta p\left(\theta \mid \alpha\right) \sum_{k=1}^{K}\left(\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \phi_{1:K}\right)\right)\left(\prod_{k=1}^{K} p\left(\phi_k \mid \beta\right)\right)\left(\prod_{i=1}^{I} p\left(y_i \mid z_{1:N}, \eta_i, \xi\right) p\left(\eta_i \mid \mu\right)\right) d\theta} \tag{3}$$

$$\int_\theta \int_{\phi_{1:K}} p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \theta, \phi, \eta, \alpha, \beta, \mu, \xi\right) d\theta d\phi_{1:K} = \int_\theta \int_{\phi_{1:K}} \prod_{k=1}^{K} p\left(\phi_k; \beta\right) \prod_{m=1}^{M}\left\{p\left(\theta_m; \alpha\right) \prod_{n=1}^{N}\left[p\left(z_{m,n} \mid \theta_m\right) p\left(w_{m,n} \mid \phi_{z_{m,n}}\right)\right] \prod_{i=1}^{I} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right\} \prod_{i=1}^{I} p\left(\eta_i; \mu\right) d\theta d\phi_{1:K} \tag{4}$$

$$p\left(\mathbf{Y}, \mathbf{w}, \mathbf{z}, \eta, \alpha, \beta, \mu, \xi\right) = \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right) \prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right] \int_{\phi_{1:K}} \prod_{k=1}^{K} p\left(\phi_k; \beta\right) \prod_{m=1}^{M} \prod_{n=1}^{N} p\left(w_{m,n} \mid \phi_{z_{m,n}}\right) d\phi_{1:K} \int_\theta \prod_{m=1}^{M} p\left(\theta_m; \alpha\right) \prod_{n=1}^{N} p\left(z_{m,n} \mid \theta_m\right) d\theta \tag{5}$$

$$= \prod_{i=1}^{I}\left[p\left(\eta_i; \mu\right) \prod_{m=1}^{M} p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V} \Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K} \Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{6}$$

$$= \prod_{i=1}^{I}\left[\mathcal{N}\left(\eta_i \mid \mu, 1\right) \prod_{m=1}^{M} h\left(y\right) \exp\left\{\left(\eta_i^T \bar{z}\right) y - A\left(\eta_i^T \bar{z}\right)\right\}\right] \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V} \Gamma\left(n_{(\bullet),v}^{k} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n_{(\bullet),v}^{k} + \beta_v\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \frac{\prod_{k=1}^{K} \Gamma\left(n_{m,(\bullet)}^{k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} n_{m,(\bullet)}^{k} + \alpha_k\right)} \tag{7}$$

2.7 Gibbs Sampling

To derive the Gibbs sampler we evaluate the individual conditional probability distributions for all unobserved variables.

2.7.1 $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu, \xi\right)$

For the purposes of sampling, we will be able to derive a representation of the joint distribution isolating a particular latent variable, z, for a word instance, n, in a document instance, m. The conditional probability with respect to this latent variable is proportional to the joint distribution up to a constant.

$$p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu, \xi\right) \propto p\left(z_{m,n}, \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu, \xi\right) \tag{8}$$

Due to the factorization of this model we can rewrite the joint distribution as the following:

$$\propto \prod_{i=1}^{I} \left[p\left(\eta_i; \mu\right) p\left(y_{m,i} \mid z_{m,1:N}, \eta_i \xi\right)\right] p\left(z_{m,n}, \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{w}, \eta, \alpha, \beta, \mu\right) \tag{9}$$

We isolate only terms that depend on $z_{m,n}$ and absorb all other constant terms into the normalization constant [10].

$$\propto \prod_{i=1}^{I} \left[p\left(y_{m,i} \mid z_{m,1:N}, \eta_i, \xi\right)\right] \left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\cdot),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V} \left(n_{(\cdot),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{10}$$

Here, $n_{m,v}^{k,-(m,n)}$ represents the count of word v in document m assigned to topic k omitting the $(m,n)^{th}$ word count. For exponential family distributions, the normalization contant, $h(y_{m,i})$, does not depend on $z_{m,n}$.

$$\propto \exp\left\{\sum_{i=1}^{I} \left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\} \left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right) \frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V} \left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{11}$$

Given this expression, $p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu\right)$ can be sampled through enumeration as seen in Equation 13. In the case of probit regression, the expression for 11 evaluates to Equation 14. Equivalently we can parameterize the model with an auxiliary variable a_i, resulting in Equation 15.

$$p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{Y}, \mathbf{w}, \eta, \alpha, \beta, \mu\right) = \frac{\exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right)\frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)}}{\sum_{k=1}^{K}\exp\left\{\sum_{i=1}^{I}\left(\eta_i^T \bar{z}\right) y_{m,i} - A\left(\eta_i^T \bar{z}\right)\right\}\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right)\frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)}} \tag{12}$$

$$p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{w}, \eta, \alpha, \beta, \mu\right) \propto \prod_{i=1}^{I}\left[\Phi\left(\eta_i^T \bar{z}\right)^{y_{m,i}}\left(1 - \Phi\left(\eta_i^T \bar{z}\right)\right)^{1-y_{m,i}}\right]\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right)\frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{13}$$

$$p\left(z_{m,n} \mid \mathbf{z}_{-(\mathbf{m},\mathbf{n})}, \mathbf{w}, \eta, \mathbf{a}, \alpha, \beta, \mu\right) \propto \prod_{i=1}^{I}\left[\delta\left(sign\left(a_{m,i}\right) = 2y_{m,i} - 1\right)\mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1\right)\right]\left(n_{m,(\bullet)}^{k,-(m,n)} + \alpha_k\right)\frac{n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,w_{m,n}}}{\sum_{v=1}^{V}\left(n_{(\bullet),w_{m,n}}^{k,-(m,n)} + \beta_{k,v}\right)} \tag{14}$$

2.7.2 $p(\eta_i \mid \mathbf{z}, \mathbf{Y}, \mu)$ or $p(\eta_i \mid \mathbf{z}, \mathbf{a}, \mu)$in the augmented probit regression model

Given that η_i and $a_{m,i}$ are distributed normally, this posterior distribution is also normal. In the case for general exponential family distributions, η_i can remain a parameter without a prior, fit with maximimum likelihood in the usual fashion.

$$p(\eta_i \mid \mathbf{z}, \mathbf{a}, \mu) = \mathcal{N}\left(\eta_i \mid \hat{\mu_i}, \hat{\mathbf{S}}_i\right) \tag{15}$$

$$\hat{\mu_i} = \hat{\mathbf{S}}_i \bar{\mathbf{Z}}^T \mathbf{a}_{(\bullet),i}$$

$$\hat{\mathbf{S}}_i^{-1} = \mathbf{I} + \bar{\mathbf{Z}}^T \bar{\mathbf{Z}}$$

2.7.3 $p(a_{m,i} \mid \mathbf{z}, \mathbf{Y}, \eta)$ in the augmented probit regression model

In the augmented probit regression model, the posterior distribution of a_i is distributed according to a truncated normal distribution.

$$p(a_{m,i} \mid \mathbf{z}, \mathbf{Y}, \eta) = truncN\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)$$

2.7.4 $p(y_{m,i} \mid \eta, \mathbf{a}, \xi)$

In our model, response variables are not always observed and are treated as latent and sampled where appropriate. There are two factors influencing predictions of the response variable, $y_{m,i}$. There is an undirected model enforcing the aforementioned constraints and providing a prior and there is the probit regression.

$$p(y_{m,i} \mid \eta, \mathbf{a}, \xi) \propto \psi(\mathbf{Y})\, \delta(sign(a_{m,i}) = y_{m,i})\, \mathcal{N}\left(a_{m,i} \mid \eta_i^T \bar{z}, 1\right) \tag{16}$$

$$p(y_{m,i} \mid \eta, \mathbf{a}, \xi) \propto \psi(\mathbf{Y})\, truncN\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right) \tag{17}$$

Again, this conditional distribution can be evaluated through enumerations and normalization.

$$p(y_{m,i} \mid \eta, \mathbf{a}, \xi) = \frac{\psi(\mathbf{Y})\, truncN\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)}{\sum_{y_{m,i}} \psi(\mathbf{Y})\, truncN\left(a_{m,i} \mid \eta_i^T \bar{z}, 1, y_{m,i}\right)} \tag{18}$$

3 Results

4 Conclusion

References

[1] International classification of disease. `http://bioportal.bioontology.org/ontologies/35686`, May 2008.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20: 121–128, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[4] D. Blumenthal. Stimulating the adoption of health information technology. *The New England Journal of Medicine*, 360(15):1477–1479, 2009.

[5] P Brown, DG Cochrane, and JR Allegra. The ngram cc classifier: A novel method of automatically creating cc classifiers based on icd9 groupings. *Advances in Disease Surveillance*, 1:30, 2006.

[6] K Crammer, M Dredze, K Ganchev, PP Talukdar, and S Carroll. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136, 2007.

[7] R Farkas and G Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(Suppl 3):S10, 2008.

[8] HR FreitasJunior, B RibeiroNeto, RF Vale, AHF Laender, and LRS Lima. Categorizationdriven cross-language retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4):501–510, 2006.

[9] I Goldstein, A Arzumtsyan, and Ö Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279, 2007.

[10] TL Griffiths and M Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[11] LV Lita, S Yu, S Niculescu, and J Bi. Large scale diagnostic code classification for medical patient records. 2008.

[12] RB Rao, S Sandilya, RS Niculescu, C Germond, and H Rao. Clinical and financial outcomes analysis with existing hospital patient records. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 416–425, 2003.

[13] B RibeiroNeto, AHF Laender, and LRS De Lima. An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5): 391–401, 2001.

[14] P Ruch, J Gobeill, I Tbahriti, and A Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *AMIA Annual Symposium Proceedings*, 2008:636, 2008.

[15] G. Surjan. Questions on validity of international classification of diseases-coded diagnoses. *International journal of medical informatics*, 54(2):77–95, 1999.