2014

Probit Normal Correlated Topic Models

Xingchen Yu

Ernest Fokoué

Probit Normal Correlated Topic Models

Xingchen Yu
Rochester Institute of Technology
98 Lomb Memorial Drive, Rochester, NY 14623, USA
xvy5021@gmail.com

Ernest Fokoué
Rochester Institute of Technology
98 Lomb Memorial Drive, Rochester, NY 14623, USA
ernest.fokoue@rit.edu

Abstract

The logistic normal distribution has recently been adapted via the transformation of multivariate Gaussian variables to model the topical distribution of documents in the presence of correlations among topics. In this paper, we propose a probit normal alternative approach to modelling correlated topical structures. Our use of the probit model in the context of topic discovery is novel, as many authors have so far concentrated solely of the logistic model partly due to the formidable inefficiency of the multinomial probit model even in the case of very small topical spaces. We herein circumvent the inefficiency of multinomial probit estimation by using an adaptation of the diagonal orthant multinomial probit in the topic models context, resulting in the ability of our topic modelling scheme to handle corpuses with a large number of latent topics. An additional and very important benefit of our method lies in the fact that unlike with the logistic normal model whose non-conjugacy leads to the need for sophisticated sampling schemes, our approach exploits the natural conjugacy inherent in the auxiliary formulation of the probit model to achieve greater simplicity. The application of our proposed scheme to a well known Associated Press corpus not only helps discover a large number of meaningful topics but also reveals the capturing of compellingly intuitive correlations among certain topics. Besides, our proposed approach lends itself to even further scalability thanks to various existing high performance algorithms and architectures capable of handling millions of documents.

Keywords: *Bayesian, Gibbs Sampler, Cumulative Distribution Function, Probit, Logit, Orthant, Efficient Sampling, Auxiliary Variable , Correlation Structure, Topic, Vocabulary, Conjugate, Dirichlet, Gaussian.*

I. Introduction

The task of recovering the latent topics underlying a given corpus of D documents has been in the forefront of active research in statistical machine learning for more than a decade, and continues to receive the dedicated contributions from many researchers from around the world. Since the introduction of Latent Dirichlet Allocation (LDA) Blei et al. (2003) and then the extension to correlated topic models (CTM) Blei and Lafferty (2006), a series of excellent contributions have been made to this exciting field, ranging from slight extension in the modelling structure to the development of scalable topic modeling algorithms capable of handling extremely large collections of documents, as well as selecting an optimal model among a collection of competing models or using the output of topic modelling as entry points (inputs) to other machine learning or data mining tasks such as image analysis and sentiment extraction, just to name a few. As far as correlated topic models are concerned, virtually all the contributors to the field have so far

concentrated solely on the use of the logistic normal topic model. The seminal paper on correlated topic modelBlei and Lafferty (2006) adopts a variational approximation approach to model fitting while subsequent authors like Mimno et al. (2008) propose a Gibbs sampling scheme with data augmentation of uniform random variables. More recently, Chen et al. (2013) presented an exact and scalable Gibbs sampling algorithm with Polya-Gamma distributed auxiliary variables which is a recent development of efficient sampling of logistic model. Despite the inseparable relationship between logistic and probit model in statistical modelling, the probit model has not yet been proposed, probably due to its computational inefficiency for multiclass classification problem and high posterior dependence between auxiliary variables and parameters. As for practical application where topic models are commonly employed, having multiple topics is extremely prevalent. In some cases, more than 1000 topics will be fitted to large datasets such as Wikipedia and Pubmed data. Therefore, using MCMC probit model in topic modeling application will be impractical and inconceivable due to its computational inefficiency. Nonetheless, a recent work on diagonal orthant probit model Johndrow et al. (2013) substantially improved the sampling efficiency while maintaining the predictive performance, which motivated us to build an alternative correlated topic modeling with probit normal topic distribution. On the other hand, probit models inherently capture a better dependency structure between topics and co-occurrence of words within a topic as it doesn't assume the IIA (independence of irrelevant alternatives) restriction of logistic models.

The rest of this paper is organized as follows: in section 2, we present a conventional formulation of topic modelling along with our general notation and the correlated topic models extension. Section 3 introduces our adaptation of the diagonal orthant probit model to topic discovery in the presence correlations among topics, along with the corresponding auxiliary variable sampling scheme for updating the probit model parameters and the remainder of all the posterior distributions of the parameters of the model. Unlike with the logistic normal formulation where the non-conjugacy leads to the need for sophisticated sampling scheme, in this section we clearly reveal the simplicity of our proposed method resulting from the natural conjugacy inherent in the auxiliary formulation of the updating of the parameters. We also show compelling computational demonstrations of the efficiency of the diagonal orthant approach compared to the traditional multinomial probit for on both the auxiliary variable sampling and the estimation of the topic distribution. Section 4 presents the performance of our proposed approach on the Associated Press data set, featuring the intuitively appealing topics discovered, along with the correlation structure among topics and the loglikelihood as a function of topical space dimension. Section 5 deals with our conclusion, discussion and elements of our future work.

II. General aspects of topic models

In a given corpus, one could imagine that each document deals with one or more topics. For instance, one of the collection considered in this paper is provided by the Associated Press and covers topics as varied as *aviation, education, weather, broadcasting, air force, navy, national security, international treaties, investing, international trade, war, courts, entertainment industry, politics*, and etc. From a statistical perspective, a topic is often modeled as a *probability distribution over words*, and as a result a given document is treated as a *mixture of probabilistic topics* Blei et al. (2003). We consider a setting where we have a total of V unique words in the reference vocabulary and K topics underlying the D documents provided. Let w_{dn} denote the n-th word in the d-th document, and let z_{dn} refer to the label of the topic assigned to the n-th word of that d-th document. Then the probability of w_{dn} is given by

$$\Pr(w_{dn}) = \sum_{k=1}^{K} \Pr(w_{dn}|z_{dn} = k) \Pr(z_{dn} = k), \tag{1}$$

where $\Pr(z_{dn} = k)$ is the probability that the nth word in the dth document is assigned to topic k. This quantity plays an important role in the analysis of correlated topic models. In the seminal article on correlated topic models Blei and Lafferty (2006), $\Pr(z_{dn} = k)$ is modeled for each document d as a function of a K-dimensional vector $\boldsymbol{\eta}_d$ of parameters. Specifically, the logistic-normal defines $\boldsymbol{\eta}_d = (\eta_d^1, \eta_d^2, \cdots, \eta_d^K)$ where the last element η_d^K is typically set to zero for identifiability and assumes with $\boldsymbol{\eta}_d \sim \texttt{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\theta_d^k = \Pr[z_{dn}^k = 1|\boldsymbol{\eta}_d] = f(\boldsymbol{\eta}_d) = \frac{e^{\eta_d^k}}{\sum_{j=1}^{K} e^{\eta_d^j}}, \quad k = 1, 2, \cdots, K-1 \quad \textit{and} \quad \theta_d^K = \frac{1}{\sum_{j=1}^{K} e^{\eta_d^j}},$$

Also, $\forall n \in \{1, 2, \cdots, N_d\}$ and $z_{dn} \sim \texttt{Mult}(\boldsymbol{\theta}_d)$, and $w_{dn} \sim \texttt{Mult}(\boldsymbol{\beta})$. With all these model components defined, the estimation task in correlated topic modelling from a Bayesian perspective can be summarized in the following posterior

$$\begin{aligned} p(\boldsymbol{\eta}_d, \boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto p(\boldsymbol{W}|\boldsymbol{Z}) \prod_{d=1}^{D} \left\{ \prod_{n=1}^{N_d} p(z_{dn}) p(\boldsymbol{\eta}_d|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\} \\ &= \prod_{k=1}^{K} \frac{\delta(C_k + \boldsymbol{\beta})}{\delta(\boldsymbol{\beta})} \prod_{d=1}^{D} \left\{ \left(\prod_{n=1}^{N_d} \theta_d^{z_{dn}} \right) \mathcal{N}(\boldsymbol{\eta}_d|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\}, \end{aligned} \tag{2}$$

where $\delta(\cdot)$ is defined using the Gamma function $Gamma(\cdot)$ so for a K-dimension vector $\boldsymbol{u}$,

$$\delta(\boldsymbol{u}) = \frac{\prod_{k=1}^{K} \Gamma(u_k)}{\Gamma\left(\sum_{k=1}^{K} u_k \right)}.$$

(3) provides the ingredients for estimating the parameter vectors $\boldsymbol{\eta}_d$ that help capture the correlations among topics, and the matrix $\boldsymbol{Z}$ that contains the topical assignments. Under the logistic normal model, sampling from the full posterior of $\boldsymbol{\eta}_d$ derived from the joint posterior in (3) requires the use of sophisticated sampling schemes like the one used in Chen et al. (2013). Although these authors managed to achieve great performances on large corpuses of documents, we thought it useful to contribute to correlated topic modelling by way of the multinomial probit. Clearly, as indicated earlier, most authors concentrated on logistic-normal even despite non-conjugacy, and the lack of probit topic modeling can be easily attributed to the inefficiency of the corresponding sampling scheme. In the most raw formulation of the multinomial probit that intends to capture the full extend of all the correlations among the topics, the topic assignment probability is defined by (3).

$$\Pr(z_{dn} = k) = \theta_d^k = \int \int \int \cdots \int \phi_K(u; \boldsymbol{\eta}_d, R) du \tag{3}$$

The practical evaluation of (3) involves a complicated high dimensional integral which is typically computationally intractable when the number of categories is greater than 4. A relaxed version of (3), one that still captures more correlation than the logit and that is also very commonly used in practice, defines θ_d^k as

$$\theta_d^k = \int_{-\infty}^{+\infty} \left\{ \prod_{j=1, j\neq k}^{K} \Phi(v + \eta_d^k - \eta_d^j) \right\} \phi(v) dv = \mathbb{E}_{\phi(v)} \left\{ \prod_{j=1, j\neq k}^{K} \Phi(V + \eta_d^k - \eta_d^j) \right\}, \tag{4}$$

where $\phi(v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}$ is the standard normal density, and $\Phi(v) = \int_{-\infty}^{v} \phi(u) du$ is the standard normal distribution function. Despite this relaxation, the multinomial probit in this formulation still has major drawbacks namely: (a) Even when one is given the vector $\boldsymbol{\eta}_d$, the calculation of θ_d^k remains computationally prohibitive even for moderate values of K. In practice, one may consider using a monte carlo approximation to that integral in (4). However, such an approach in the context of a large corpus with many underlying latent topics renders the probit formulation almost unusable. (b) As far as the estimation of $\boldsymbol{\eta}_d$ is concerned, a natural approach to sampling from the posterior of $\boldsymbol{\eta}_d$ in this context would be to use the Metropolis-Hastings updating scheme, since the full posterior in this case is not available. Unfortunately, the Metropolis in this case is excruciatingly slow with poor mixing rates and high sensitivity to the proposal distribution. It turns out that an apparently appealing solution in this case could come from the auxiliary variable formulation as described in Albert and Chib (1993). Unfortunately, even this promising formulation fails catastrophically for moderate values K as we will demonstrate in the subsequent section, due to the high dependency structure between auxiliary variables and parameters. Essentially, the need for Metropolis is avoided by defining an auxiliary vector Y_d of dimension K. For $n = 1, \cdots, N_d$, we consider the vector z_{dn} containing the current topic allocation and we repeatedly sample Y_{dn} from a K-dimensional multivariate Gaussian until the component of Y_{dn} that corresponds to the non-zero index in z_{dn} is the largest of all the components of Y_{dn}, ie.

$$Y_{dn}^{z_{dn}} = \max_{k=1,\cdots,K} \{Y_{dn}^k\}. \tag{5}$$

The condition in (5) typically fails to be fulfilled even when K is moderately large. In fact, we demonstrate later that in some cases, it becomes impossible to find a vector Y_{dn} satisfying that condition. Besides, the dependency of Y_{dn} on the current value of $\boldsymbol{\eta}_d$ further complicates the sampling scheme especially in the case of large topical space. In the next section, we remedy these inefficiencies by proposing and developing our adaptation of the diagonal orthant multinomial probit.

III. Diagonal Orthant Probit for Correlated Topic Models

In a recent work, Johndrow et al. (2013) developed the diagonal orthant probit approach to multicategorical classification. Their approach circumvents the bottlenecks mentioned earlier and substantially improves the sampling efficiency while maintaining the predictive performance. Essentially, the diagonal orthant probit approach successfully makes the most of the benefits of binary classification, thereby substantially reducing the high dependency that made the condition (5) computationally unattainable. Indeed, with the diagonal orthant multinomial model, we achieved three main benefits

- A more tractable and easily computatble definition of topic distribution $\theta_d^k = \Pr(z_{dn} = k | \boldsymbol{\eta}_d)$

- A clear and very straightforward and adaptable auxiliary variable sampling scheme
- The capacity to handle a very large number of topics due to the efficiency and low dependency.

Under the diagonal orthant probit model, we have

$$\theta_d^k = \frac{(1-\Phi(-\eta_d^k))\prod_{j\neq k}\Phi(-\eta_d^j)}{\sum_{\ell=1}^{K}(1-\Phi(-\eta_d^\ell))\prod_{j\neq \ell}\Phi(-\eta_d^j)}. \tag{6}$$

The generative process of our probit normal topic models is essentially identical to logistic topic models except that the topic distribution for each document now is obtained by a probit transformation of a multivariate Gaussian variable (6). As such, the generating process of a document of length N_d is as follows:

1. Draw $\eta \sim \texttt{MVN}(\mu, \Sigma)$ and transform η_d into topic distribution θ_d where each element of θ is computed as follows:

$$\theta_d^k = \frac{(1-\Phi(-\eta_d^k))\prod_{j\neq k}\Phi(-\eta_d^j)}{\sum_{\ell=1}^{K}(1-\Phi(-\eta_d))\prod_{j\neq \ell}\Phi(-\eta_d^j)}. \tag{7}$$

2. For each word position $n \in (1, \cdots, N_d)$
   (a) Draw a topic assignment $Z_n \sim \texttt{Mult}(\theta_d)$
   (b) Draw a word $W_n \sim \texttt{Mult}(\varphi^{z_n})$

Where $\Phi(\cdot)$ represents the cumulative distribution of the standard normal. We specify a Gaussian prior for $\boldsymbol{\eta}_d$, namely $(\boldsymbol{\eta}_d|\cdots) \sim \mathcal{N}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Throughout this paper, we'll use $\phi_K(\cdot)$ to denote the K-dimensional multivariate Gaussian density function,

$$\phi_K(\boldsymbol{\eta}_d; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^K|\Sigma|}}\exp\left\{-\frac{1}{2}(\boldsymbol{\eta}_d-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta}_d-\boldsymbol{\mu})\right\}.$$

To complete the Bayesian analysis of our probit normal topic model, we need to sample from the joint posterior

$$p(\boldsymbol{\eta}_d, \boldsymbol{Z}_d|\boldsymbol{W}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto p(\boldsymbol{\eta}_d|\boldsymbol{\mu}, \boldsymbol{\Sigma})p(\boldsymbol{Z}_d|\boldsymbol{\eta}_d)p(\boldsymbol{W}|\boldsymbol{Z}_d). \tag{8}$$

As noted earlier, the second benefit of the diagonal orthant probit model lies in its clear, simple, straightforward yet powerful auxiliary variable sampling scheme. We take advantage of that diagonal orthant property when dealing with the full posterior for $\boldsymbol{\eta}_d$ given by

$$p(\boldsymbol{\eta}_d|\boldsymbol{W}, \boldsymbol{Z}_d, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto p(\boldsymbol{\eta}_d|\boldsymbol{\mu}, \boldsymbol{\Sigma})p(\boldsymbol{Z}_d|\boldsymbol{\eta}_d). \tag{9}$$

While sampling directly from (9) is impractical, defining a collection of auxiliary variables $\boldsymbol{Y}_d$ allows a scheme that samples from the joint posterior $p(\boldsymbol{\eta}_d, \boldsymbol{Z}_d, \boldsymbol{Y}_d|\boldsymbol{W}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ using the following:

For each document d, the matrix $\boldsymbol{Y}_d \in \mathbb{R}^{N_d \times K}$ contains all the values of the auxiliary variables,

$$\boldsymbol{Y}_d = \begin{bmatrix} Y_{d1}^1 & Y_{d1}^2 & \cdots & Y_{d1}^k & \cdots & Y_{d1}^K \\ Y_{d2}^1 & Y_{d2}^2 & \cdots & Y_{d2}^k & \cdots & Y_{d2}^K \\ \vdots & \vdots & \cdots & \ddots & \cdots & \vdots \\ Y_{d,N_d-1}^1 & Y_{d,N_d-1}^2 & \cdots & Y_{d,N_d-1}^k & \cdots & Y_{d,N_d-1}^K \\ Y_{d,N_d}^1 & Y_{d,N_d}^2 & \cdots & Y_{d,N_d}^k & \cdots & Y_{d,N_d}^K \end{bmatrix}$$

Each row $Y_{dn} = (Y_{dn}^1, \cdots, Y_{dn}^k, \cdots, Y_{dn}^K)^\top$ of $\boldsymbol{Y}_d$ has K components, and the diagonal orthant updates them readily using the following straightforward sampling scheme: Let k be the current topic allocation for the nth word.

- For the component of Y_{dn} whose index corresponds to the label of current topic assignment of word n sample from a truncated normal distribution with variance 1 restricted to positive outcomes
$$(Y_{dn}^k | \eta_d^k) \sim \mathcal{N}_+(\eta_d^k, 1) \quad z_{dn}^k = 1$$
- For all components of Y_{dn} whose indices do correspond to the label of current topic assignment of word n sample from a truncated normal distribution with variance 1 restricted to negative outcomes
$$(Y_{dn}^j | \eta_d^j) \sim \mathcal{N}_-(\eta_d^j, 1) \qquad z_{dn}^j \neq 1$$

Once the matrix $\boldsymbol{Y}_d$ is obtained, the sampling scheme updates the parameter vector $\boldsymbol{\eta}_d$ by conveniently drawing

$$(\boldsymbol{\eta}_d | \boldsymbol{Y}_d, \boldsymbol{A}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim MVN(\boldsymbol{\mu}_{\eta_d}, \boldsymbol{\Sigma}_{\eta_d}),$$

where

$$\boldsymbol{\mu}_{\eta_d} = \boldsymbol{\Sigma}_{\eta_d}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{X}_d^\top \boldsymbol{A}^{-1}\text{vec}(\boldsymbol{Y}_d)) \quad \text{and} \quad \boldsymbol{\Sigma}_{\eta_d} = (\boldsymbol{\Sigma}^{-1} + \boldsymbol{X}_d^\top \boldsymbol{A}^{-1}\boldsymbol{X}_d)^{-1}.$$

with $\boldsymbol{X}_d = \boldsymbol{1}_{N_d} \otimes \boldsymbol{I}_K$ and $\text{vec}(\boldsymbol{Y}_d)$ representing the row-wise vectorization of the matrix $\boldsymbol{Y}_d$. Adopting the fully Bayesian treatment of our probit normal correlated topic model, we add an extra layer to the hierarchy in order to capture the variation in the mean vector and the variance-covariance matrix of the parameter vector $\boldsymbol{\eta}_d$. Taking advantage of conjugacy, we specify a normal-Inverse-Wishart prior for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, namely,

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = NIW(\boldsymbol{\mu}_0, \kappa_0, \Psi_0, \nu_0),$$

meaning that $\boldsymbol{\Sigma}|\nu_0, \Psi_0 \sim IW(\Psi_0, \nu_0)$ and $(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}, \kappa_0) \sim MVN(\boldsymbol{\mu}_0, \Sigma/\kappa_0)$. The corresponding posterior is normal-inverse-Wishart, so that we can write

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\eta}) = NIW(\boldsymbol{\mu}', \kappa', \Psi', \nu'),$$

where $\kappa' = \kappa_0 + D$, $\nu' = \nu_0 + D$, $\boldsymbol{\mu}' = \frac{D}{D+\kappa_0}\bar{\boldsymbol{\eta}} + \frac{\kappa_0}{D+\kappa_0}\boldsymbol{\mu}_0$, and

$$\Psi' = \Psi_0 + Q + \frac{\kappa_0}{\kappa_0 + D}(\bar{\boldsymbol{\eta}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{\eta}} - \boldsymbol{\mu}_0)^\top,$$

where

$$Q = \sum_{d=1}^{D} (\boldsymbol{\eta}_d - \bar{\boldsymbol{\eta}})(\boldsymbol{\eta}_d - \bar{\boldsymbol{\eta}})^\top.$$

As far as sampling from the full posterior distribution of Z_{dn} is concerned, we use the expression

$$\Pr[z_{dn}^k = 1 | \boldsymbol{Z}_{\neg n}, w_{dn}, \boldsymbol{W}_{\neg dn}] \propto p(w_{dn} | z_{dn}^k = 1, \boldsymbol{W}_{\neg dn}, \boldsymbol{Z}_{\neg n}) \theta_d^k \propto \frac{C_{k,\neg n}^{w_{dn}} + \beta_{w_{dn}}}{\sum_{j=1}^V C_{k,\neg n}^j + \sum_{j=1}^V \beta_j} \theta_d^k.$$

where the use of $C_{\cdot,\neg n}$ is used to indicate that the nth is not included in the topic or document under consideration.

IV. Computational results on the Associated Press data

In this section, we used a famous Associated Press data set from GrÃijn and Hornik (GrÃijn and Hornik) in R to uncover the word topic distribution, the correlation structure between various topics as well as selecting optimal models. The Associated Press corpus consists of 2244 documents and 10473 words. After preprocessing the corpus by picking frequent and common terms, we reduced the size of the words from 10473 to 2643 for efficient sampling.

In our first experimentation, we built a correlated topic modelling structure based on the traditional multinomial probit and then tested the computational speed for key sampling tasks. The high posterior dependency structure between auxiliary variables and parameters make multinormal probit essentially unscalable for situations where it is impossible for the sampler to yield a random variate of the auxiliary variable corresponding the current topic allocation label that is also the maximum (5). For a random initialization of topic assignment, the sampling of auxiliary variable cannot even complete one single iteration. In the case of good initialization of topical prior $\boldsymbol{\eta}_d$ which leads to smooth sampling of auxiliary variables, the computational efficiency is still undesirable and we observed that for larger topical space such as K=40, the auxiliary variable stumbled again after some amount of iterations, indicating even good initialization will not ease the troublesome dependency relationship between the auxiliary variables and parameters in larger topical space. Unlike with the traditional probit model for which the computation of θ_d^k is virtually impractical for large K,the diagonal orthant approach makes this computation substantially faster ever for large K. The comparison of the computational speed of two essential sampling tasks between the multinomial probit model and digonal orthant probit model are shown as below in table 1 (1).

In addition to the drastic improvement of the overall sampling efficiency, we noticed that the computational complexity for sampling the auxiliary variable and topic distribution is close to O(1) and O(K) respectively, suggesting that probit normal topic model now becomes an attainable and feasible tool of the traditional correlated topic model.

Central to topic modelling is the need to determine for a given corpus the optimal number of latent topics. As it is the case for most latent variable models, this task can be formidable at times, and there is no consensus among machine learning researchers as to which of the existing methods is the best. Figure (1) shows the loglikelihood as a function of the number of topics discovered in the model. Apart from the loglikelihood, many other techniques are commonly used such as perplexity, harmonic mean method and so on.

As we see, the optimal number of topics in this case is 30. In table (2), we show a subset of the 30 topics uncovered where each topic is represented by the 10 most frequent words. It can be seen that our probit normal topic model is able to capture the co-occurrence of words within topics

Sampling Task (K=10)	MNP	DO Probit
Topic Distribution θ	18.3	0.06
Auxiliary variable Y_d	(108 to NA)	3.09
Sampling Task (K=20)	MNP	DO Probit
Topic Distribution θ	63	0.13
Auxiliary variable Y_d	(334 to NA)	3.39
Sampling Task (K=30)	MNP	DO Probit
Topic Distribution θ	123	0.21
Auxiliary variable Y_d	(528 to NA)	3.49
Sampling Task (K=40)	MNP	DO Probit
Topic Distribution θ	211.49	0.33
Auxiliary variable Y_d	(1785 to NA)	3.79

Table 1: *All the numbers in this table represent the processing time (in seconds), and are computed in R on PC using a parallel algorithm acting on 4 CPU cores. NA here represents situations where it is impossible for the sampler to yield a random variate of the auxiliary variable corresponding the current topic allocation label that is also the maximum*

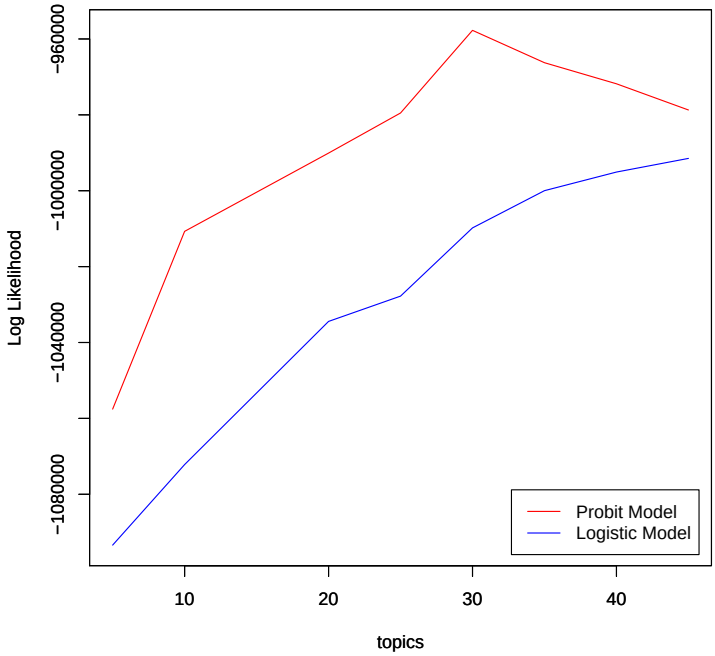

Figure 1: *Loglikelihood as a function of the number of topics*

successfully. In figure 2, we also show the correlation structure between various topics which is the essential purpose of employing the correlated topic model. Evidently, the correlation captured intuitively reflect the natural relationship between similar topics.

	Topic 25	**Topic 18**	**Topic 23**	**Topic 11**	**Topic 1**	**Topic 24**	**Topic 27**
Word1	court	company	bush	students	tax	fire	air
Word2	trial	billion	senate	school	budget	water	plane
Word3	judge	inc	vote	meese	billion	rain	flight
Word4	prison	corp	dukakis	student	bill	northern	airlines
Word5	convicted	percent	percent	schools	percent	southern	pilots
Word6	jury	stock	bill	teachers	senate	inches	aircraft
Word7	drug	workers	kennedy	board	income	fair	planes
Word8	guilty	contract	sales	education	legislation	degrees	airline
Word9	fbi	companies	bentsen	teacher	taxes	snow	eastern
Word10	sentence	offer	ticket	tax	bush	temperatures	airport

	Topic 6	**Topic 12**	**Topic 20**	**Topic 2**	**Topic 22**	**Topic 16**	**Topic 15**
Word1	percent	space	military	soviet	aid	police	dollar
Word2	stock	shuttle	china	gorbachev	rebels	arrested	yen
Word3	index	soviet	chinese	bush	contras	shot	rates
Word4	billion	nasa	soldiers	reagan	nicaragua	shooting	bid
Word5	prices	launch	troops	moscow	contra	injured	prices
Word6	rose	mission	saudi	summit	sandinista	car	price
Word7	stocks	earth	trade	soviets	military	officers	london
Word8	average	north	rebels	treaty	ortega	bus	gold
Word9	points	korean	hong	europe	sandinistas	killing	percent
Word10	shares	south	army	germany	rebel	arrest	trading

	Topic19	**Topic 14**	**Topic 7**	**Topic 4**	**Topic 30**	**Topic 8**	**Topic 17**
Word1	iraq	trade	israel	navy	percent	south	film
Word2	kuwait	percent	israeli	ship	oil	africa	movie
Word3	iraqi	farmers	jewish	coast	prices	african	music
Word4	german	farm	palestinian	island	price	black	theater
Word5	gulf	billion	arab	boat	cents	church	actor
Word6	germany	japan	palestinians	ships	gasoline	pope	actress
Word7	saudi	agriculture	army	earthquake	average	mandela	award
Word8	iran	japanese	occupied	sea	offers	blacks	band
Word9	bush	tons	students	scale	gold	apartheid	book
Word10	military	drought	gaza	guard	crude	catholic	films

Table 2: *Representation of topics discovered by our method*

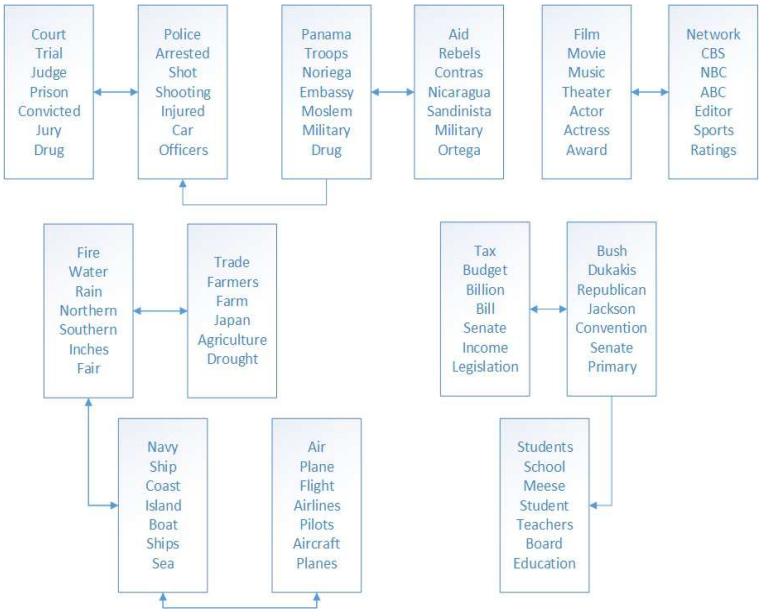

Figure 2: *Graphical representation of the correlation among topics*

V. Conclusion and Discussion

In the context of topic modelling where many other researchers seem to have avoided it. By adapting the diagonal orthant probit model, we proposed a probit alternative to the logit approach to the topic modeling. Compared to the multinomial probit model we constructed, our topic discovery scheme using diagonal orthant probit model enjoyed several desirable properties; First,we gained the efficiency in computing the topic distribution $\boldsymbol{\theta}_d^k$; Second, we achieved a clear and very straightforward and adaptable auxiliary variable sampling scheme that substantially reduced the strength of the dependence structure between auxiliary variables and model parameters, responsible for absorbing state in the Markov chain; Thirdly, as a consequence of good mixing, our approach made the probit model a viable and competitive alternatives to its logistic counterpart. In addition to all these benefits, our proposed method offers a straightforward and inherent conjugacy, which helps avoid those complicated sampling schemes employed in the logistics normal probit model.

In the Associated Press example explored in the previous section, not only does our method produce a better likelihood than the logistic normal topic model with variational EM, but also discovers meaningful topics along with underlying correlation structure between topics. Overall, the method we developed in this paper offers another feasible alternatives in the context of correlated topic model that we hope will be further explored and extended by many other researchers

Based on the promising results we have seen in this paper, the probit normal topic model opens the door for various future works. For instance, Salomatin et al. (2009) proposed a multi-field correlated topic model by relaxing the assumption of using common set of topics globally among all documents, which can also be applied to the probit model to enrich the comprehensiveness of structural relationships between topics . Another potential direction would be to enhance the scalability of the model. Currently we used a simple distributed algorithm proposed by Yao et al.

(2009) and Newman et al. (2009) for efficient Gibbs sampling. The architecture for topic models presented by Smola and Narayanamurthy (2010) can be further utilized to reduce the computational complexity substantially while delivering comparable performance. Furthermore, a novel sampling method involving the Gibbs Max-Margin Topic Zhu et al. (2013) will further improve the computational efficiency.

References

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*(422), 669–679.

Blei, D. M. and J. D. Lafferty (2006). Correlated topic models. In *In Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120. MIT Press.

Blei, D. M., A. Y. Ng, M. I. Jordan, and J. Lafferty (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 2003.

Chen, J., J. Zhu, Z. Wang, X. Zheng, and B. Zhang (2013). Scalable inference for logistic-normal topic models. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*, pp. 2445–2453. Curran Associates, Inc.

GrÃijn, B. and K. Hornik. topicmodels: An r package for fitting topic models. *Journal of Statistical Software 40*(13), 1–30.

Johndrow, J., K. Lum, and D. B. Dunson (2013). Diagonal orthant multinomial probit models. In *JMLR Proceedings*, Volume 31 of *AISTATS*, pp. 29–38. JMLR.

Mimno, D., H. M. Wallach, and A. Mccallum (2008). Gibbs sampling for logistic normal topic models with graph-based priors.

Newman, D., A. Asuncion, P. Smyth, and M. Welling (2009, December). Distributed algorithms for topic models. *J. Mach. Learn. Res. 10*, 1801–1828.

Salomatin, K., Y. Yang, and A. Lad (2009). Multi-field correlated topic modeling. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA*, pp. 628–637.

Smola, A. and S. Narayanamurthy (2010). An architecture for parallel topic models. In *VLDB*.

Yao, L., D. Mimno, and A. McCallum (2009). Efficient methods for topic model inference on streaming document collections. In *KDD*.

Zhu, J., N. Chen, H. Perkins, and B. Zhang (2013). Gibbs max-margin topic models with data augmentation. *CoRR abs/1310.2816*.