Topics Over Nonparametric Time: A Supervised Topic Model Using Bayesian Nonparametric Density Estimation

Daniel D. Walker, Kevin Seppi, and Eric K. Ringger
Computer Science Department
Brigham Young University
Provo, UT, 84604
danw@lkers.org, {kseppi, ringger}@cs.byu.edu

Abstract

We propose a new supervised topic model that uses a nonparametric density estimator to model the distribution of real-valued metadata given a topic. The model is similar to Topics Over Time, but replaces the beta distributions used in that model with a Dirichlet process mixture of normals. The use of a nonparametric density estimator allows for the fitting of a greater class of metadata densities. We compare our model with existing supervised topic models in terms of prediction and show that it is capable of discovering complex metadata distributions in both synthetic and real data.

1 Introduction

Supervised topic models are a class of topic models that, in addition to modeling documents as mixtures of topics, each with a distribution over words, also model metadata associated with each document. Document collections often include such metadata. For example, timestamps are commonly associated with documents that represent the time of the document's creation. In the case of online product reviews, "star" ratings frequently accompany written reviews to quantify the sentiment of the review's author.

There are three basic reasons that make supervised topic models attractive tools for use with document collections that include metadata. *Better Topics*: one assumption that is often true for document collections is that the topics being discussed are correlated with information that is not necessarily directly encoded in the text. Using the metadata in the inference of topics provides an extra source of information, which could lead to an improvement in modeling the topics that are found. *Prediction*: given a trained supervised topic model and a new document with missing metadata, one can predict the value of the metadata variable for that document. Even though timestamps are typically included in modern, natively digital, documents they may be unavailable or wrong for historical documents that have dbeen igitized using OCR. Also, even relatively modern documents can have missing or incorrect timestamps due to user error or system mis-configuration. For example, in the full Enron e-mail corpus[1], there are 793 email messages with a timestamp before 1985, the year Enron was founded. Of these messages 271 have a timestamp before the year 100. *Analysis*: in order to understand a document collection better, it is often helpful to understand how the metadata and topics are related. For example, one might want to analyze the development of a topic over time, or investigate what the presence of a particular topic means in terms of the sentiment being expressed by the author. One may, for example, plot the distribution of the metadata given a topic from a trained model. Several supervised topic models can be found in the literature and will be discussed in more detail in Section 3. These models make assumptions about the way in which the metadata are distributed given the topic or require the user to specify their own assumptions. Usually, this approach involves using a unimodal distribution, and the same distribution family is used to model the metadata across all topics.

[1] http://www.cs.cmu.edu/~enron

These modeling assumptions are problematic. First, it is easy to imagine metadata and topics that have complex, multi-modal relationships. For example, the U.S. has been involved in two large conflicts with Iraq over the last 20 years. A good topic model trained on news text for that period should ideally discover an Iraq topic and successfully capture the bimodal distribution of that topic in time. Existing supervised topic models, however, will either group both modes into a single mode, or split the two modes into two separate topics. Second, it seems incorrect to assume that the metadata will be distributed similarly across all topics. Some topics may remain fairly uniform over a long period of time, others appear quickly and then fade out over long periods of time (e.g., terrorism after 9/11), others enter the discourse gradually over time (e.g., healthcare reform), still others appear and disappear in a relatively short period of time (e.g., many political scandals).

To address these issues, we introduce a new supervised topic model, Topics Over Nonparametric Time (TONPT), based on the Topics Over Time (TOT) model [12]. Where TOT uses a per-topic beta distribution to model topic-conditional metadata distributions, TONPT uses a nonparametric density estimator, a Dirichlet process mixture (DPM) of normals.

The remainder of the paper is organized as follows: in Section 2 we provide a brief discussion of the Dirichlet process and show how a DPM of normals can be used to approximate a wide variety of densities. Section 3 outlines related work. In Section 4 we introduce the TONPT model and describe the collapsed Gibbs sampler we used to efficiently conduct inference in the model on a given dataset. Section 5 describes experiments that were run in order to compare TONPT with two other supervised topic models and a baseline. Finally, in Section 6 we summarize our results and contributions.

2 Estimating Densities with Dirichlet Process Mixtures

Significant work has been done in the document modeling community to make use of Dirichlet process mixtures with the goal of eliminating the need to specify the number of components in a mixture model. For example, it is possible to cluster documents without specifying a-priori the number of clusters by replacing the Dirichlet-multinomial mixing distribution in the Mixture of Multinomials document model with a Chinese Restaurant Process. The CRP is the distribution over partitions created by the clustering effect of the Dirichlet process [1]. So, one way of using the Dirichlet process is in model-based clustering applications where it is desirable to let the number of clusters be determined dynamically by the data, instead of being specified by the user.

The DP is a distribution over probability measures G with two parameters: a base measure G_0 and a total mass parameter m. Random probability measures drawn from a DP are generally not suitable as likelihoods for continuous random variates because they are discrete. This complication can be overcome by convolving the G with a continuous kernel density f [9, 5, 6]:

$$\begin{aligned} G &\sim \boldsymbol{DP}(m, G_0) \\ x_i|G &\sim \int f(x_i|\theta) dG(\theta) \end{aligned}$$

This model is equivalent to an infinite mixture of f distributions with hierarchical formulation:

$$\begin{aligned} G &\sim \boldsymbol{DP}(m, G_0) \\ \theta_i|G &\sim G \\ x_i|\theta_i &\sim f(x_i|\theta) \end{aligned}$$

In our work we use the normal distribution for f. The normal distribution has many advantages that make it a useful choice here. First, the parameters map intuitively to the idea that the θ parameters in the DPM are the "locations" of the point masses of G and so are a natural fit for the mean parameter of the normal distribution. Second, because the normal is conjugate to the mean of a normal with known variance, we can also choose a conjugate G_0 that has intuitive parameters and simple posterior and marginal forms. Third, the normal is almost trivially extensible to multivariate cases. Fourth, the normal can be centered anywhere on the positive or negative side of the origin which is not true, for example, of the gamma and beta distributions. Finally, just as any 1D signal can be approximated with a sum of sine waves, almost any probability distribution can be approximated with a weighted sum of normal densities.

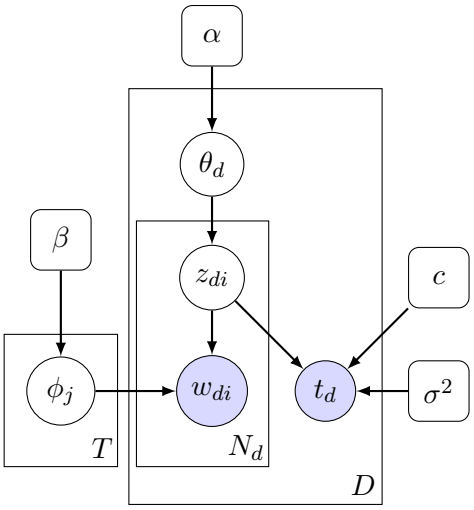


Figure 1: The Supervised LDA model.

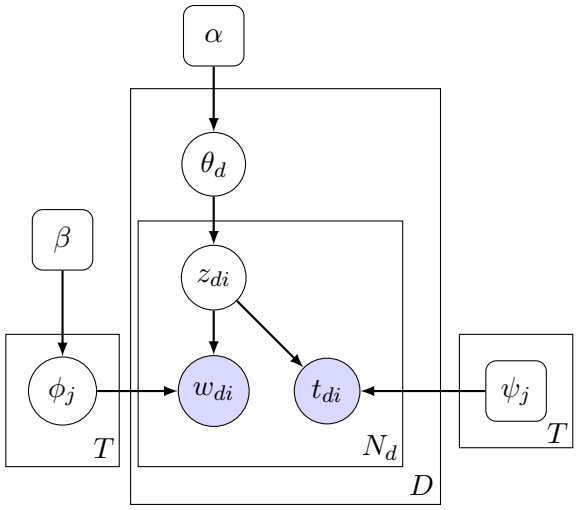


Figure 2: The Topics Over Time model.

3 Related Work

In this section we will describe the three models which are most closely related to our work. In particular, we focus on the issues of prediction and the posterior analysis of metadata distributions in order to highlight the strengths and weaknesses of each model.

The most closely related models to TONPT are Supervised LDA (sLDA) [3] and Topics Over Time [12]. sLDA uses a generalized linear model (GLM) to regress the metadata given the topic proportions of each document. GLMs are flexible in that they allow for the specification of a link and a dispersion function that can change the behavior of the regression model. In practice, however, making such a change to the model requires non-trivial modifications to the inference procedure used to learn the topics and regression co-efficients. In the original sLDA paper, an identity link function and normal dispersion distribution were used. The model, shown in Figure 1, has per-document timestamp variables $t_d \sim \textit{Normal}(c \cdot \overline{z_d}, \sigma^2)$, where c is the vector of linear model coefficients and $\overline{z_d}$ is a topic proportion vector for document d (See Table 1 for a discription of the other variables in the models shown here). This configuration leads to a stochastic EM inference procedure in which one alternately samples from the complete conditional for each topic assignment, given the current values of all the other variables, and then finds the regression coefficients that minimize the sum squared residual of the linear prediction model. Variations of sLDA have been used successfully in several applications including modeling the voting patterns of U.S. legislators [7] and links between documents [4].

Prediction in sLDA is very straightforward, as the latent metadata variable for a document can be marginalized out to produce a vanilla LDA complete conditional distribution for the topic assignments. The procedure for prediction can thus be as simple as first sampling the topic assignments for each word in an unseen document given the assignments in the training set, and then taking the dot product between the estimated topic proportions for the document and the GLM coefficients. In terms of the representation of the distribution of metadata given topics, however, the model is somewhat lacking. The coefficients learned during inference convey only one-dimensional information about the correlation between topics and the metadata. A large positive coefficient for a given topic indicates that documents with a higher proportion of that topic tend to have higher metadata values, and a large negative coefficient means that documents with a higher proportion of that topic tend to have lower metadata values. Coefficients close to zero indicate low correlation between the topic and the metadata.

In TOT, metadata are treated as per-word observations, instead of as a single per-document observation. The model, shown in Figure 2, assumes that each per-word metadatum t_{di} is drawn from a per-topic beta distribution: $t_{di} \sim \textit{Beta}(\psi_{z_{di}1}, \psi_{z_{di}2})$. The inference procedure for TOT is a stochastic EM algorithm, where the topic assignments for each word

are first sampled with a collapsed Gibbs sampler and then the shape parameters for the per-topic beta distributions are point estimated using the Method of Moments based on the mean and variance of the metadata values for the words assigned to each topic.

Prediction in TOT is not as straightforward as for sLDA. Like sLDA, it is possible to integrate out the random variables directly related to the metadata and estimate a topic distribution for a held-out document using vanilla LDA inference. However, because the model does not include a document-level metadata variable, there is no obvious way to predict a single metadata value for held-out documents. We describe a prediction procedure in Section 5, based on work one by Wang and McCallum, that yields acceptable results in practice.

Despite having a more complicated prediction procedure, TOT yields a much richer picture of the trends present in the data. It is possible with TOT, for example, to get an idea of not only whether the metadata are correlated with a topic, but also to see the mean and variance of the per-topic metadata distributions and even to show whether the distribution is skewed or symmetric.

Another related model is the Dirichlet Multinomial Regression (DMR) model [11]. Whereas the sLDA and TOT models both model the metadata generatively, i.e., as random variables conditioned on the topic assignments for a document, the DMR forgoes modeling the metadata explicitly, putting the metadata variables at the "root" of the graphical model and conditioning the document distributions over topics on the metadata values. By forgoing a direct modeling of the metadata, the DMR is able to take advantage of a wide range of metadata types and even to include multiple metadata measurements (or "features") per document. The authors show how, conditioning on the metadata, the DMR is able to outperform other supervised topic models in terms of its ability to fit the observed words of held-out documents, yielding lower perplexity values. The DMR is thus able to accomplish one of the goals of supervised topic modeling very well (the increase in topic quality). However, because it does not propose any distribution over metadata values, it is difficult to conduct the types of analyses or missing metadata predictions possible in TOT and sLDA without resorting to ad-hoc procedures. Because of these deficiencies, we leave the DMR out of the remaining discussion of supervised topic models.

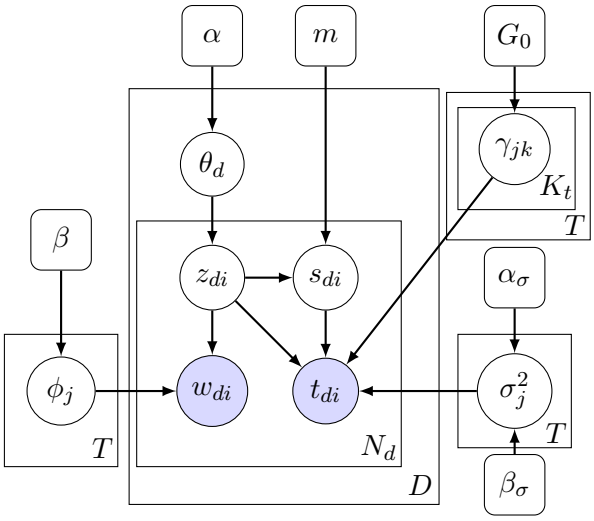


Figure 3: TONPT as used in sampling.

4 Topics Over Nonparametric Time

TONPT models metadata variables associated with each word in the corpus as being drawn from a topic-specific Dirichlet process mixture of normals. In addition, TONPT employs a common base measure G_0 for all of the per-topic DPMs, for which we use a normal with mean μ_0 and variance σ_0^2.

The random variables are distributed as follows:

$$\begin{aligned}
\theta_d|\alpha &\sim \textit{Dirichlet}(\alpha) \\
\phi_t|\beta &\sim \textit{Dirichlet}(\beta) \\
z_{di}|\theta &\sim \textit{Categorical}(\theta_d) \\
w_{di}|z_{di},\phi &\sim \textit{Categorical}(\phi_{z_{di}}) \\
\sigma_j^2|\alpha_\sigma,\beta_\sigma &\sim \textit{InverseGamma}(\alpha_\sigma,\beta_\sigma) \\
G_j|G_0,m &\sim \textit{DP}(G_0,m) \\
t_{di}|G_{z_{di}},\sigma_{z_{di}}^2 &\sim \int f(t_{di};\gamma,\sigma_{z_{di}}^2)dG_{z_{di}}(\gamma)
\end{aligned}$$

where $f(\cdot;\gamma,\sigma^2)$ is the normal p.d.f. with mean γ and variance σ^2. Also, $j \in \{1,\ldots,T\}$, $d \in \{1,\ldots,D\}$, and, given a value for d, $i \in \{1,\ldots,N_d\}$. We note that, as in TOT, the fact that the metadata variable is repeated per-word leads to a deficient generative model, because the metadata are typically observed at a document level and the assumed constraint that all of the metadata values for the words in a document be equivalent is not modeled. The advantage of

Symbol	Meaning
Common Supervised Topic Modeling Variables	
α	Prior parameter for document-topic distributions
θ_d	Parameter for topic mixing distribution for document d
β	Prior parameter for the topic-word distributions
ϕ_j	Parameter for the jth topic-word distribution
z_{di}	Topic label for word i in document d
$\mathbf{z}_{-di}$	All topic assignments except that for z_{di}
$\mathbf{w}$	Vector of all word token types
w_{di}	Type of word token i in document d
t_{di}	Timestamp for word i in document d
t_d	Timestamp for document d
$\mathbf{t}$	Vector of all metadata variable values
$\hat{t}$	A predicted value for the metadata variable
D	The number of documents
T	The number of topics
V	The number of word types
N_d	The number of tokens in document d
TONPT Specific Variables	
m	Total mass parameter for DP mixtures
s_{di}	DP component membership for word i in document d
$\mathbf{s}_{-di}$	All DP component assignments except that for s_{di}
G_0	The base measure of the DP mixtures
μ_0	The mean of the base measure
σ_0^2	The variance of the base measure
γ_{jk}	The mean of the kth mixture component for topic j
$\boldsymbol{\gamma}$	A vector of all the γ values
$\boldsymbol{\gamma}_{-jk}$	$\boldsymbol{\gamma}$ without γ_{jk}
σ_j^2	The variance of the components of the jth DP mixture
$\boldsymbol{\sigma}^2$	A vector of all the DPM σ^2s
$\alpha_\sigma, \beta_\sigma$	Shape and scale parameters for prior on topic σs
K_j	The number of unique observed γs for topic j
n_j	The number of tokens assigned to topic j
n_{jk}	The number of tokens assigned to the kth component of topic j
n_{dj}	The number of tokens in document d assigned to topic j
n_{jv}	The number of times a token of type v was assigned to topic j
$K_{z_{di}}^{<di}$	The number of unique γs observed for topic z_{di} before the ith token of document d
$\tau^{(jk)}$	The set of all t_{di} s.t. $z_{di} = j$ and $s_{di} = k$
$f(y; \mu, \sigma^2)$	The normal p.d.f. at y with mean μ and variance σ^2

Table 1: Mathematical symbols used in the models and derivations of this paper. The common symbols are shared by TONPT, sLDA, and TOT.

this approach is that this configuration simplifies inference, and also naturally balances the plurality of the word variables with the singularity of the metadata variable, allowing the metadata to exert a similarly scaled influence on the topic assignments during inference. In addition, this modeling choice allows for a more fine-grained labeling of documents (e.g., at the word, phrase, or paragraph level) and for finer grained prediction. For example, while timestamps should probably be the same for all words in a document, sentiment does not need to meet this constraint–there are often positive comments even in very negative reviews.

This model does not lend itself well to inference and sampling because of the integral in the distribution over t_{di}. A typical modification that is made to facilitate sampling in mixture models is to use an equivalent hierarchical model. Another modification that is typically made when sampling in mixture models is to separate the "clustering," or mixing, portion of the distribution from the prior over mixture component parameters. The mixing distribution in a DPM is the distribution known as the Chinese Restaurant Process. The Chinese Restaurant Process is used to select an assignment to one of the points that makes up the DP point process for each data observation drawn from G. The locations of these points are independently drawn from G_0.

Figure 3 shows the model that results from decomposing the Dirichlet process into these two component pieces. The K_j unique γ values that have been sampled so far for each topic j are drawn from G_0. The s_{di} variables are indicator variables that take on values in $1, \ldots, K_j$ and represent which of the DPM components each t_{di} was drawn from. This model has the following changes to the variable distributions:

$$s_{di}|z_{di}, \mathbf{s}^{<di}, m \sim \begin{cases} = k \text{ with prob } \propto n_{z_{di},k}^{<di} \\ \quad \text{for } k = 1, \ldots, K_{z_{di}}^{<di} \\ = K_{z_{di}}^{<di} + 1 \text{ with prob } \propto m \end{cases}$$

$$\gamma_{jk}|G_0 \sim G_0$$

$$t_{di}|z_{di}, s_{z_{di}}, \boldsymbol{\gamma}, \boldsymbol{\sigma}^2 \sim f\left(t_{di}; \gamma_{z_{di}s_{di}}, \sigma_{z_{di}}^2\right)$$

Where $\mathbf{s}^{<di}$ refers to all the $s_{d'i'}$ that came "before" s_{di} and before is defined to mean all (d', i') such that $(d' < d)$ or $(d' = d \text{ and } i' < i)$. Likewise, $n_{z_{di},k}^{<di}$ is the count of the number of times that $s_{d'i'} = k$ for all d', i' before s_{di} and $K_{z_{di}}^{<di}$ is the highest value of any

$s_{d'i'}$ (number of unique observed γs) before s_{di}. So, conditioned on z_{di} the s_{di} are distributed according to a Chinese Restaurant Process with mass parameter m.

The θ and ϕ variables in the model are nuisance variables: they are not necessary for the assignment of tokens to topics or for the estimation of the distributions of the response variables so, as is typical when conducting Gibbs sampling on these models, we integrate them out before sampling.

4.1 Gibbs Sampler Conditionals

Now we derive the complete conditionals for the collapsed Gibbs sampler used for inference in the model. There are four groups of variables that must be sampled during inference: the per-word topic labels z, the per word DPM component assignment variables s, the DPM component means $\boldsymbol{\gamma}$, and the per-topic DPM component variances $\boldsymbol{\sigma}^2$. Note that, because of normal-normal conjugacy, it would be possible to collapse the γ variables from the model. We choose to sample values for γ anyway because the parameters of the DPM are useful artifacts in their own right, as they enable rich posterior analyses of the per-topic metadata distributions.

4.1.1 Complete Conditional for z and s

We choose to sample z_{di} and s_{di} in a block, since the calculations necessary to sample z_{di} include those sufficient to sample both variables jointly.

$$\begin{aligned}[z_{di}, s_{di}] =& p(z_{di} = j, s_{di} = k | \mathbf{z}_{-di}, \mathbf{s}_{-di}, \boldsymbol{\sigma}^2, \mathbf{w}, \quad (1)\\ & \mathbf{t}, m, \boldsymbol{\gamma}, \alpha_\sigma, \beta_\sigma, G_0, \alpha, \beta)\\ \propto& \alpha_{\star dj} \frac{\beta_{\star j w_{di}}}{\sum_{v=1}^{V} \beta_{\star jv}} \quad (2)\\ & \cdot \begin{cases} \frac{n_{jk}}{n_j + m} f(t_{di}; \gamma_{jk}, \sigma_j^2) \\ \quad \text{if } k \leq |\gamma_j|, \\ \frac{m}{n_j + m} f(t_{di}; \mu_0, \sigma_0^2 + \sigma_j^2) \\ \quad \text{if } k = |\gamma_j| + 1 \end{cases}\end{aligned}$$

where $\alpha_{\star dj} = \alpha_j + n_{dj}$, $\beta_{\star jv} = \beta_v + n_{jv}$.

4.1.2 Complete Conditional for γ

When sampling a z_{di}, s_{di} pair if $s_{di} = K_{z_{di}} + 1$ (i.e., we are creating a new component for the DPM for that topic), then we need to draw a new γ for the z_{di}th DPM. Also, each γ_{jk} needs to be resampled each iteration of the Gibbs sampler.

Let $\tau^{(jk)} = \{t_{di} : z_{di} = j \text{ and } s_{di} = k\}$ ordered arbitrarily, which groups the t_{di} by the topic and DPM component that they are associated with. The complete conditional for each γ is:

$$\begin{aligned}[\gamma_{jk}] =& p(\gamma_{jk} | \mathbf{s}, \mathbf{t}, \mathbf{w}, \mathbf{z}, \boldsymbol{\gamma}_{-jk}, \boldsymbol{\sigma}^2, \quad (3)\\ & \alpha_\sigma, \beta_\sigma, m, \alpha, \beta, \mu_0, \sigma_0^2)\\ =& f\left(\gamma_{jk}; \mu_{jk\star}, \sigma^2_{jk\star}\right) \quad (4)\end{aligned}$$

where $\sigma_\star^2 = \left(\frac{1}{\sigma_0^2} + \frac{|\tau^{(jk)}|}{\sigma_j^2}\right)^{-1}$ and

$\mu_\star = \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{|\tau^{(jk)}|} \tau_i^{(jk)}}{\sigma_j^2}\right) \cdot \sigma_\star^2$

4.1.3 Complete Conditional for σ^2

The complete conditional is a common result for gamma-normal conjugacy. In this case, the likelihood is restricted to those t_{di} for which $z_{di} = j$:

$$\left[\sigma_j^2\right] = \textit{InverseGamma}\left(\alpha_{\sigma_\star}, \beta_{\sigma_\star}\right) \quad (5)$$

where $\alpha_{\sigma_\star} = \alpha_\sigma + \frac{n_j}{2}$,

$$\beta_{\sigma_\star} = \beta_\sigma + \frac{\sum_{d=1}^{D} \sum_{i=1}^{N_d} [\mathbf{1}_j(z_{di})(\gamma_{z_{di}, s_{di}} - t_{di})^2]}{2},$$

and $\mathbf{1}_j(x)$ is the Kronecker delta.

5 Experiments

We inferred topic assignments and metadata distributions for several real-world datasets using sLDA, TOT, TONPT, and a baseline method that we will refer to as PostHoc in which a vanilla LDA model is inferred over the dataset and then a linear model is fit to the metadata using the document topic proportions as predictors.

Because it is difficult to know a-priori what form the distributions over metadata given topics will take in real-world data, we also ran one experiment with synthetic data, where the metadata distributions were prespecified. Synthetic data was used in order to determine whether TONPT can accurately recover complex metadata distributions in conjunction with topic distributions.

The focus of our experiments was to measure quantitatively how well each model can predict metadata values on unseen data and to assess qualitatively (e.g., via inspection) whether the trained models capture human intuition and domain knowledge with respect to the correlations between topics and metadata values.

5.1 Data

We ran our experiments on three real-world datasets. For each dataset the timestamps of the documents were extracted and used as the metadata. For all real-world datasets, stopwords were removed using the stopwords file included in the MALLET topic modeling toolkit [10]. In addition, words that occurred in more than a half of the documents in a dataset and those that occurred in fewer than 1% were culled. Words were converted to lowercase, and documents that were empty after pre-processing were removed. Finally, only for the TOT model, the metadata were all normalized to the $(0, 1)$ interval to accommodate usage of the beta distribution.

The first dataset consists of the State of the Union Addresses delivered by Presidents of the United States from the first address by George Washington in 1790 to the second address by Barack Obama in 2010. The data was prepared in the manner similar to that of Wang and McCallum [12], in which addresses were subdivided into individual documents by paragraph, resulting in 7507 (three-paragraph) documents. The metadata for this dataset were address timestamps which were normalized to the interval $[0, 1]$.

The second dataset was the LDC-annotated portion of the Enron corpus [2]. This dataset consists of 4,935 emails, that were made public as part of the investigation of illegal activities by the Enron Corporation. It covers approximately one year of time (January 2001 through December 2001). The metadata timestamps for this dataset were extracted from the *Date* header field of the e-mail messages.

The final dataset was the Reuters 21578 corpus[8]. We used the subset of the articles for which topical tags are available, which consists of 11,367 documents. The articles were written during a time interval that spans most of the year 1987. The documents were processed using the same feature selection as for the other two datasets with an additional step in which variants of the names of the months were removed. These words are especially common in this corpus (e.g., in datelines) and provide a strong signal that is not based on the topical content of the articles (i.e., they allowed the models to "cheat"). After feature selection there were 10,230 non-empty documents in the final dataset.

5.2 Procedure

In our prediction experiments, models were trained on 90% of the documents and then were used to predict the metadata values for the remaining 10%. This was repeated in a cross-validation scheme ten times, with the training and evaluation sets being randomly sampled each time. Prediction quality was evaluated using the formula for the coefficient of determination (R^2) used by Blei and McAuliff [3]:

$$R^2(\mathbf{t}, \hat{\mathbf{t}}) = 1 - \frac{\sum_d (t_d - \hat{t}_d)^2}{\sum_d (t_d - \bar{t})^2},$$

where t_d is the actual metadatum for document d, $\hat{t}_d$ is the prediction and $\bar{t}$ is the mean of the observed t_ds. For linear models this metric measures the proportion of the variability in the data that is accounted for by the model. More generally, it is one minus the relative efficiency of the supervised topic model predictor to a predictor that always predicts the mean of the observed data points. This value can be negative in cases where the model being evaluated performs worse than the mean predictor. The means and standard errors of the R^2 values across all ten folds were recorded. In order to assess the statistical significance of the results, a one-sided permutation test was used to calculate p-values for the hypothesis that the mean R^2 for the model with the highest mean R^2 was greater than the mean R^2 for each of the other models being tested. P-values less than 0.05 were considered significant.

As discussed above, prediction in the case of sLDA is quite simple. In the case of TOT and TONPT, prediction is complicated by the fact that these models have per-word metadata variables, and not per-document variables. In addition, they do not produce a prediction using a simple dot product, but instead they provide a distribution over predicted values given a topic assignment. In order to perform prediction in TOT one finds the metadata value with maximal posterior probability given the topic assignments for all of the

words in the test document[12]:

$$\hat{t}_d = \arg\max_t \prod_i p(t|z_{di})$$

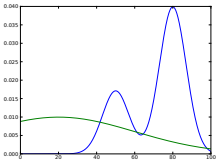
In order to approximate this value, we first infer topic assignments for each word in the document using a version of the model in which the metadata and related variables have been integrated out (i.e., vanilla LDA). Next, because the posterior is a fairly complicated product, and difficult to maximize directly, we approximate by choosing several discrete points and check the value of the posterior at each test point. In the original TOT paper, the candidate points were chosen to represent decades. In an attempt to be more general and to choose candidates that are likely to be of high posterior probability we generate candidates by sampling a metadata value for each word from the beta distribution for the topic assigned to that word. The mean of the sampled points is also added as a candidate. Finally, to generate a prediction the posterior density is calculated at each of the candidates and the one producing the greatest value is chosen.

We found that in the case of TONPT the multimodality of the $p(t|z_{di})$ distribution caused this prediction algorithm to perform poorly. For TONPT, predictions are determined by first estimating the θ_d parameter for the test document using samples obtained from the model with the metadata marginalized out, and then using θ_d to estimate the mean of $p(t|z_{di})$ as the θ_d weighted average of the means of the topic DPMs.

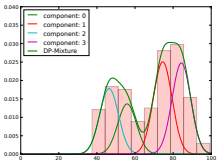
For the TONPT runs, G_0 was chosen to be a normal with mean and variance equal to the sample mean and variance for the observed metadata, α_σ was 2.0, β_σ was 1.0, and m was 1. For all runs, the document-topic parameter $\alpha = 0.1$, and the topic-word parameter $\beta = 0.01$.

5.3 Synthetic Data Results

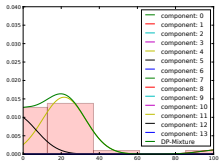
The synthetic dataset was created such that there are 2 topics and a vocabulary of 5 words: "common", "semicommon1", "semicommon2", "rare1" and "rare2". The "common" word occurs with 0.6 probability in both topics, "semicommon1" is slightly more likely than "semicommon2" in the first topic, and slightly less likely in the second topic. The "rare1" word is much more likely in the first topic than the second and "rare2" is much more likely in the second topic than the first, but both are much less common in general than the "semicommon"s.



(a) "True" metadata distributions.



(b) Distribution learned for Topic 0



(c) Distribution learned for Topic 1

Figure 4: The estimated metadata distributions discovered for the synthetic dataset.

Each topic was given a fixed metadata distribution:

$$t_0 \sim 0.3 \cdot f(50, 7) + 0.7 \cdot f(80, 7)$$
$$t_1 \sim f(20, 40)$$

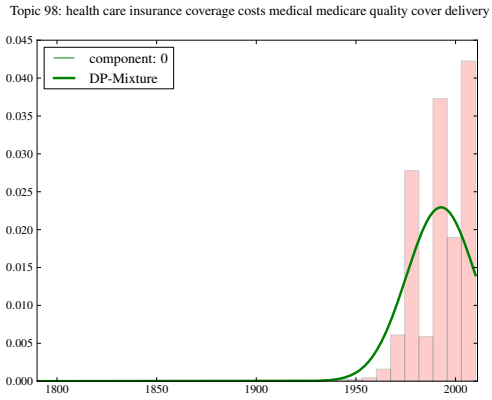
Figure 4 shows how, for one run of the inference procedure, the model was able to separate the two topics and recreate the original metadata distributions with high degree of fidelity. Some runs result in slightly better approximations, while others do worse, but these plots seem to be representative of TONPT's performance on this task.

5.4 Prediction Results

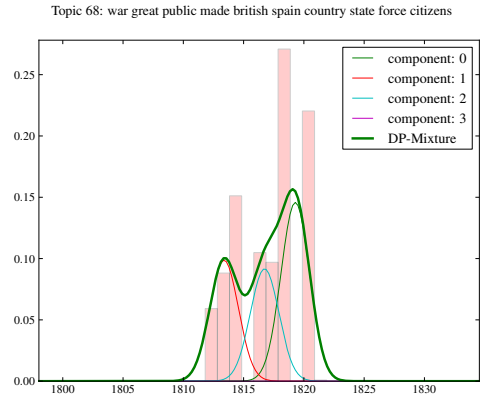
Table 2 shows the performance of the various models for the prediction task with 40 topics, which we found to be a number of topics at which peak performance was observed for most of the models. It can be seen that TONPT is significantly superior on the State of the Union and Reuters data, though TOT does come out ahead on the Enron dataset (but not significantly so).

5.5 Posterior Analysis

Figure 5 shows the distribution over time for two topics found during runs of TONPT on the State of the Union dataset. The topic shown in 5a is typical of



(a) Time distribution for a health care topic



(b) Time distribution for early 1800s conflicts topic

Figure 5: Ratings distributions for two topics found in different runs of TONPT.

Data	Model	Mean R^2	Std Err	p-val
SotU	PostHoc	0.8099	0.0053	0.004
	sLDA	0.8180	0.0046	0.029
	TOT	0.6945	0.0073	0.000
	TONPT	**0.8306**	0.0035	N/A
Enron	PostHoc	0.2434	0.0092	0.002
	sLDA	0.2638	0.0141	0.026
	TOT	**0.3137**	0.0179	N/A
	TONPT	**0.2836**	0.0175	0.137
Reuters	PostHoc	0.1031	0.0072	0.006
	sLDA	0.0775	0.0132	0.010
	TOT	-0.7873	0.0447	0.004
	TONPT	**0.1948**	0.0312	N/A

Table 2: Prediction results for the 3 real-world datasets. Values that are not statistically significantly different from the best are highlighted. P-values are from a 1-sided permutation test against the results from the model with the highest mean R^2.

the majority of the distributions we find (a DPM with only one observed component and thus very close to a simple symmetric distribution). It shows the relatively recent rise in prevalence of the topic of health care in U.S. politics.

The topic shown in 5b is an example of a more complex distribution. This particular example appears to capture several conflicts the United States was involved in during the early 1800s, including The War of 1812 and several conflicts related to the Seminole Wars in Florida (which was a Spanish territory until it was ceded to the U.S. in 1821).

6 Conclusion and Future Work

We have presented TONPT, a supervised topic model that models metadata using a nonparametric density estimator. The model accomplishes the goal of accommodating a wider range of metadata distributions and, in the case of the datasets that we evaluated against, prediction performance remains competitive with previous models. Future work could extend the model to multivariate metadata, such as temporal-spatial data including both timestamps and geolocation information. For example, a multidimensional version of TONPT could be used to capture the development of trends in Twitter data, identifying areas where topics originate and how they spread across the country over time. A multivariate normal component distribution would also capture correlations between metadata elements through a topic covariance matrix.

References

[1] David Aldous, Illdar Ibragimov, Jean Jacod, and David Aldous. Exchangeability and related topics. In *cole d't de Probabilits de Saint-Flour XIII 1983*, volume 1117 of *Lecture Notes in Mathematics*, pages 1–198. Springer Berlin / Heidelberg, 1985. 10.1007/BFb0099421.

[2] Michael W. Berry, Murray Brown, and Ben Signer. 2001 topic annotated Enron email data set, 2007.

[3] David M. Blei and Jon D. McAuliffe. Supervised topic models. In *arXiv:1003.0783*, March 2010.

[4] J. Chang and D. Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, pages 81–88, 2009.

[5] M.D. Escobar. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, pages 268–277, 1994.

[6] M.D. Escobar and M. West. Computing nonparametric hierarchical models. *Practical nonparametric and semiparametric Bayesian statistics*, pages 1–22, 1998.

[7] S.M. Gerrish and D.M. Blei. Predicting legislative roll calls from text. In *Proceedings of the 28th Annual International Conference on Machine Learning*, number 11, 2011.

[8] D. Lewis. Reuters-21578 text categorization test collection. `http://www.research.att.com/~lewis`, 1997.

[9] A.Y. Lo. On a class of bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, 12(1):351–357, 1984.

[10] Andrew Kachites McCallum. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`, 2002.

[11] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence*, pages 411–418, 2008.

[12] Xuerui Wang and Andrew McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference(KDD'06)*, Philadelphia, PA, August 2006.