Accounting for Burstiness in Topic Models

Gabriel Doyle GDOYLE@LING.UCSD.EDU
Department of Linguistics, University of California, San Diego, La Jolla CA 92093-0108, USA

Charles Elkan ELKAN@CS.UCSD.EDU
Dept. of Computer Science and Engineering, University of California, San Diego, La Jolla CA 92093-0404, USA

Abstract

Many different topic models have been used successfully for a variety of applications. However, even state-of-the-art topic models suffer from the important flaw that they do not capture the tendency of words to appear in bursts; it is a fundamental property of language that if a word is used once in a document, it is more likely to be used again. We introduce a topic model that uses Dirichlet compound multinomial (DCM) distributions to model this burstiness phenomenon. On both text and non-text datasets, the new model achieves better held-out likelihood than standard latent Dirichlet allocation (LDA). It is straightforward to incorporate the DCM extension into topic models that are more complex than LDA.

1. Introduction

The effectiveness of a topic model is dependent on the appropriateness of its generative process for the task at hand. For most common tasks, any computationally feasible generative model will be a substantial simplification of the true generative process. Nevertheless, some tractable generative models are more reflective of the true generative process than others. In this paper, we propose a new generative process for topic models that significantly improves the statistical fidelity of the process with minimal additional model complexity. Specifically, we replace the multinomial distributions in standard latent Dirichlet allocation (LDA) (Blei et al., 2003) by Dirichlet compound multinomial (DCM) distributions (Madsen et al., 2005; Elkan, 2006). The result is a better model for text data and for at least some other non-text data.

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

Our primary concern in the current study is accounting for the phenomenon of burstiness. Church and Gale (1995) note that real texts systematically exhibit this phenomenon: a word is more likely to occur again in a document if it has already appeared in the document. Importantly, the burstiness of a word and its semantic content are positively correlated; words that are more informative are also more bursty. The multinomial distribution does not take burstiness into account (Rennie et al., 2003, Sect. 4.1), so it gives an inaccurate model for the distribution of words in texts.

The phenomenon of burstiness is not limited to text. In Section 4 we present an example of bursty data in the financial realm. Burstiness also intuitively occurs in other types of data that have been modeled using topic models, including gene expression and computer vision data (Airoldi et al., 2007; Fei-Fei & Perona, 2005). If a gene is transcribed once in a cell, then it is more likely to be transcribed again. And if a patch with certain properties occurs once in an image, then it is more likely that similar patches will occur again.

The new DCMLDA model is only slightly more complex than standard LDA. As a result, the LDA component in complex topic models, such as Pachinko allocation (Li & McCallum, 2006) and correlated topic models (Blei & Lafferty, 2005), can be replaced with a DCMLDA component. This should enable those models to account for burstiness and thereby improve their effectiveness.

Because it uses DCMs to represent topics, the DCMLDA model can capture the tendency of the same topic to manifest itself with different words in different documents. Suppose that there is a natural "sports" topic in a corpus, with the words "rugby" and "hockey" being equally common overall. Within a document, though, one appearance of "rugby" makes a second appearance of "rugby" more likely than a first appearance of "hockey." The DCM distributions in DCMLDA can represent this fact, while a standard LDA model cannot. This property allows a single DCMLDA topic to explain related aspects of docu-

ments more effectively than a single LDA topic. Thus, we hypothesize, a DCMLDA model with a few topics can fit a corpus as well as an LDA model with many topics. This hypothesis is confirmed by the experimental results below.

2. Overview of Models

The DCMLDA model combines the DCM and LDA models, gaining the advantages of each. We review the two component models before discussing DCMLDA.

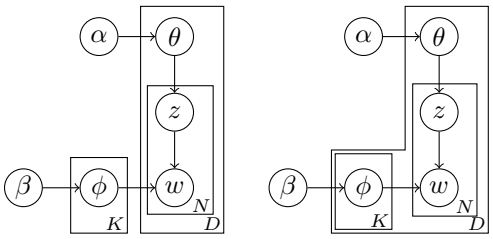
Latent Dirichlet allocation (LDA). LDA has been discussed in detail elsewhere (Blei et al., 2001; Blei et al., 2003; Griffiths et al., 2004; Heinrich, 2005), so we present only an overview here. The LDA generative model notionally posits that an author generates a document in two steps. First, the author determines the probability of each topic in the document. Each topic is a multinomial distribution over words, so to choose a word the author first draws a topic and then draws a word based on that topic. The graphical model for LDA is shown in Figure 1(a), with the unobserved variables distributed as follows:

$$\theta \sim \text{Dirichlet}(\alpha) \qquad z \sim \text{Multinomial}(\theta)$$
$$\phi \sim \text{Dirichlet}(\beta) \qquad w \sim \text{Multinomial}(\phi).$$

This generative process does not account for burstiness of words. The only way that burstiness can manifest itself is indirectly, as a consequence of how topics are distributed. The fact that a document contains the word "rugby" from a sports topic, for instance, makes it more likely that the document contains other words from the same sports topic. Thus, the document is likely to contain a second instance of the word "rugby." However, because the sports topic is the same across the corpus, the presence of any sports word in a document will have a similar effect. That is, an appearance of the word "rugby" also indirectly makes an appearance of the word "hockey" more likely, which is not a desirable phenomenon.

The LDA model is bursty in topics, even though it is not in words: the presence of one word from a given topic in a document makes other words in the document more likely to be generated by the same topic. However, because the LDA generative process does not account for word-level burstiness, LDA may in fact be excessively bursty at the topic level. The reason is that each occurrence of a word is treated as independent extra evidence for its topic.

An LDA model has two Dirichlet hyperparameters, α and β, which condition θ and ϕ respectively. Different values for the hyperparameters cause different inferred



(a) LDA (b) DCMLDA

Figure 1. Alternative graphical models.

values of ϕ and θ. In general α and β are vectors that can be learned (Blei et al., 2003; Fei-Fei & Perona, 2005). However, often they are kept fixed and uniform, meaning that each vector component is set to the same scalar value.

Learning the hyperparameters can provide information about the corpus: α indicates how semantically diverse documents are, with lower α indicating increased diversity, while β indicates how similar the topics are, with higher β indicating more similarity between topics. Learning non-uniform values for the hyperparameters allows different words and topics to have different tendencies; some topics can be more general than others (e.g., function words versus medical jargon), and some words can be likely to appear in more topics than others (e.g., words with multiple senses).

Despite not accounting for burstiness, LDA is an effective model that has proven useful for its ability to model documents as varying mixtures of shared topics. From a trained LDA model, one can infer the multinomial distributions θ that give the probability of each topic in each document. These distributions can then be used for many tasks, including classifying new documents and measuring similarity between documents.

Dirichlet compound multinomial (DCM). The DCM model (Madsen et al., 2005) captures burstiness, but it has no notion of topic. DCM uses a bag-of-bags-of-words generative process. In this process, each document is formed by drawing a document-specific multinomial distribution ϕ from a shared Dirichlet distribution, and then drawing words w according to ϕ. In the DCM model, each document is composed of words drawn from a single multinomial. This multinomial can be viewed as a document-specific subtopic, or aspect, of the high-level topic β. The β vector is the only parameter of DCM, so unlike the hyperparameters in LDA, it must be non-uniform.

Since topics are drawn from a Dirichlet distribution in LDA also, it is perhaps not immediately obvious why DCM accounts for the burstiness of words and LDA does not. The answer lies in the 1:1 mapping

between subtopics and documents in the DCM model. In LDA, the multinomial distribution of words in each topic depends on the whole corpus, but DCM multinomial distributions are document-specific.

Turning to the mathematics of the models, a key difference is that multinomial parameters are constrained to sum to one, unlike Dirichlet parameters. This gives the DCM model one extra degree of freedom to represent a topic. By working out an exponential-family approximation of the DCM, Elkan (2006) shows explicitly that this degree of freedom allows the DCM to discount multiple observations of the same word. In bursty texts, additional appearances of a word are less surprising than its first appearance. The smaller the sum of the Dirichlet parameters β, the more the emission of words is bursty. As the Dirichlet parameters tend to infinity, a DCM distribution approaches equivalence with a multinomial distribution.

A single DCM model represents one high-level topic that has alternative aspects. It cannot represent multiple distinct topics. Because of the 1:1 mapping between multinomials and documents, in a DCM model each document comes entirely from one subtopic. All these subtopics are closely related because the sum of the β vector is typically quite high (a few hundred). Elkan (2006) extended the DCM model to a mixture of DCM distributions. This model can be trained to represent a set of documents where each document comes from a different high-level topic, but it cannot represent the scenario where a single document contains words from more than one high-level topic.

DCMLDA. To combine the advantages of DCM and LDA, we need a model that allows multiple topics in a single document, while still making the topics document-specific to account for burstiness. Figure 1 contrasts the LDA and DCMLDA graphical models, while Algorithm 1 is the DCMLDA generative process.

In LDA, for each topic k, one multinomial distribution ϕ_k is drawn from Dirichlet(β) and is used in all documents. In DCMLDA, for each topic k and each document d a fresh multinomial word distribution ϕ_{kd} is drawn. Each topic k has a different, non-uniform β_k vector. For each document d, ϕ_{kd} is drawn according to Dirichlet(β_k), so the instances of each topic are linked across documents. Having per-document instances of each topic allows for variations in the probability of each word in the same topic in different documents, which is the phenomenon of burstiness.

Algorithm 1 DCMLDA Generative Model

for document $d \in \{1, \ldots, D\}$ **do**
  draw topic distribution $\theta_d \sim \mathrm{Dir}(\alpha)$
  for topic $k \in \{1, \ldots, K\}$ **do**
    draw topic-word distribution $\phi_{kd} \sim \mathrm{Dir}(\beta_k)$
  end for
  for word $n \in \{1, \ldots, N_d\}$ **do**
    draw topic $z_{dn} \sim \theta_d$
    draw word $w_{dn} \sim \phi_{z_{dn}d}$
  end for
end for

The change from a single set of multinomial topics to multiple sets of multinomial subtopics shifts the focus of attention in DCMLDA modeling. Let V be the size of the vocabulary, let K be the number of topics, and let D be the number of documents in the corpus. In LDA, ϕ is the focus, a $V \times K$ array of word probabilities given topics. In DCMLDA, ϕ is three-dimensional ($V \times K \times D$), measuring word likelihoods for each topic, for each document. Since in DCMLDA ϕ depends on the specific document, it is not a representation of the data that has sharply reduced dimensionality. Instead, with DCMLDA the focus of attention is β, which is a two-dimensional array of Dirichlet parameters for words given topics. As mentioned in the previous section, the β values are not constrained to sum to one. This gives DCMLDA an extra K degrees of freedom that allow it to capture word-level burstiness within each topic. The β values have a similar intuitive interpretation to the ϕ values in LDA. In particular, higher β values mean that a word is more likely in a given topic. Thus one can still use β values to identify the most common words in each topic.

3. Methods of Inference

Both the standard LDA model and the DCMLDA model have five unobserved variables: α, β, ϕ, θ, and z. These variables can be classified into two groups: the per-document or per-word parameters ϕ, θ, and z, and the hyperparameters α and β. Given a training set of documents, we learn appropriate values for the variables by alternating between optimizing the topic parameters given the hyperparameters, and optimizing the hyperparameters given the topic parameters. Neither of these optimizations can be done analytically, but both yield to known estimation procedures. Specifically, for fixed values of the α vector and β array, we do collapsed Gibbs sampling to find the distribution of z given the documents. If desired, ϕ and θ can be computed straightforwardly from samples of z. Given a z sample, values of α and β that maximize the likelihood of the training documents are obtained by Monte Carlo expectation-maximization.

In this and subsequent sections, the notation $\beta_{\cdot k}$ in-

dicates that β is a two-dimensional array, with one column for each topic k, so $\beta_{\cdot k}$ is what was informally called β_k previously. Similarly, the notation $\alpha_{\cdot}$ is used to emphasize that α is a vector.

Gibbs sampling. Gibbs sampling for DCMLDA is similar to the method for LDA, which Heinrich (2005) explains in detail. We present a condensed derivation, highlighting what is novel for DCMLDA sampling. We start by factoring the complete likelihood of the model: $p(w, z|\alpha_{\cdot}, \beta_{\cdot\cdot}) = p(w|z, \beta_{\cdot\cdot})p(z|\alpha_{\cdot})$. The first probability is an average over all possible ϕ distributions:

$$\begin{aligned} p(w|z, \beta_{\cdot\cdot}) &= \int_{\phi} p(z|\phi)p(\phi|\beta_{\cdot\cdot})d\phi \\ &= \int_{\phi} p(\phi|\beta_{\cdot\cdot}) \prod_{d} \prod_{n=1}^{N_d} \phi_{w_{dn} z_{dn} d}\, d\phi \\ &= \int_{\phi} p(\phi|\beta_{\cdot\cdot}) \prod_{d,k,t} (\phi_{tkd})^{n_{tkd}} d\phi. \end{aligned}$$

Expanding $p(\phi|\beta_{\cdot\cdot})$ as a Dirichlet distribution yields

$$\begin{aligned} p(w|z, \beta_{\cdot\cdot}) &= \int_{\phi} \left[\prod_{d,k} \frac{1}{B(\beta_{\cdot k})} \prod_{t} (\phi_{tkd})^{\beta_{tk}-1} \right] \\ &\quad \times \left[\prod_{d,k,t} (\phi_{tkd})^{n_{tkd}} \right] d\phi \\ &= \prod_{d,k} \int_{\phi} \prod_{t} (\phi_{tkd})^{\beta_{tk}-1+n_{tkd}} d\phi \\ &= \prod_{d,k} \frac{B(n_{\cdot kd} + \beta_{\cdot k})}{B(\beta_{\cdot k})}. \end{aligned} \tag{1}$$

Above, $B(\cdot)$ is the multidimensional Beta function, and n_{tkd} is how many times word t is assigned topic k in document d. DCMLDA and LDA are structurally identical over the α-to-z pathway, so $p(z|\alpha_{\cdot})$ in DCMLDA is the same as for LDA:

$$p(z|\alpha_{\cdot}) = \prod_{d} \frac{B(n_{\cdot\cdot d} + \alpha_{\cdot})}{B(\alpha_{\cdot})}. \tag{2}$$

Combining Equations 1 and 2 yields that the complete likelihood $p(w, z|\alpha_{\cdot}, \beta_{\cdot\cdot})$ is

$$\prod_{d} \left[\frac{B(n_{\cdot\cdot d} + \alpha_{\cdot})}{B(\alpha_{\cdot})} \prod_{k} \frac{B(n_{\cdot kd} + \beta_{\cdot k})}{B(\beta_{\cdot k})} \right]. \tag{3}$$

To perform collapsed Gibbs sampling, we need to calculate $p(z_i|z_{-i}, w)$, where z_{-i} is the set of topic assignments to all words but w_i. Letting n_{tkd} be the count of word t in topic k and document d in the complete corpus $\{w_{-i} \cup w_i\}$, and letting n'_{tkd} be the count for the limited corpus w_{-i}, we get the DCMLDA Gibbs sampling equation:

$$\begin{aligned} p(z_i|z_{-i}, w) &= \frac{p(z, w)}{p(z_{-i}, w)} \\ &= \frac{B(n_{\cdot\cdot d_i} + \alpha_{\cdot})B(n_{\cdot z_i d_i} + \beta_{\cdot z_i})}{B(n'_{\cdot\cdot d_i} + \alpha_{\cdot})B(n'_{\cdot z_i d_i} + \beta_{\cdot z_i})} \\ &= \frac{(n_{\cdot z_i d_i} + \alpha_{z_i} - 1)(n_{w_i z_i d_i} + \beta_{w_i z_i} - 1)}{(\sum_k n_{\cdot k d_i} + \alpha_k - 1)(\sum_t n_{t z_i d_i} + \beta_{t z_i} - 1)}. \end{aligned}$$

Hyperparameter EM. Many applications of LDA are successful using default uniform values for α and β, for example $\alpha = 50/K$ and $\beta = .01$, where K is the number of topics, as suggested by Griffiths and Steyvers (2004). Therefore it is not always necessary to learn the hyperparameters in LDA. However, it is imperative to learn the hyperparameters in DCMLDA. The information contained in the ϕ values with LDA is contained in the β values with the DCMLDA model.

Ideally, we would compute optimal α and β values by maximizing the likelihood $p(w|\alpha, \beta)$ directly. Unfortunately, even evaluating this likelihood is intractable. What can be computed is the complete likelihood $p(w, z|\alpha, \beta)$. Based on this, we use single-sample Monte Carlo EM to learn α and β. The single-sample method is recommended by Celeux et al. (1996) because it is computationally simple and generally outperforms multiple-sample Monte Carlo EM. Algorithm 2 summarizes the method as applied to DCMLDA.

To implement the M-step of the algorithm we need to find α and β that maximize Equation 3, given the current topic assignments. Expanding the Beta functions yields

$$\begin{aligned} p(w, z|\alpha, \beta) = &\prod_{d} \left[\frac{(\prod_k \Gamma(n_{\cdot kd} + \alpha_k))\Gamma(\sum_k \alpha_k)}{(\prod_k \Gamma(\alpha_k))\Gamma(\sum_k n_{\cdot kd} + \alpha_k)} \right] \\ &\times \prod_{d,k} \left[\frac{(\prod_t \Gamma(n_{tkd} + \beta_{wk})\Gamma(\sum_t \beta_{tk})}{(\prod_t \Gamma(\beta_{tk}))\Gamma(\sum_t n_{tkd} + \beta_{tk})} \right]. \end{aligned}$$

Algorithm 2 Single-Sample Monte Carlo EM

Start with initial $\alpha_{\cdot}$ and $\beta_{\cdot\cdot}$.
repeat
  Run Gibbs sampling to steady-state
  Choose a specific topic assignment for each word using Gibbs sampling
  Choose $\alpha_{\cdot}$ and $\beta_{\cdot\cdot}$ to maximize complete likelihood $p(w, z|\alpha_{\cdot}, \beta_{\cdot\cdot})$
until convergence of $\alpha_{\cdot}$ and $\beta_{\cdot\cdot}$

Now we convert to log-likelihood:

$$L(\alpha,\beta;w,z) = \sum_{d,k}[\log\Gamma(n_{\cdot kd}+\alpha_k) - \log\Gamma(\alpha_k)] + \sum_{d}[\log\Gamma(\sum_k \alpha_k) - \log\Gamma(\sum_k n_{\cdot kd}+\alpha_k)] + \sum_{d,k,t}[\log\Gamma(n_{tkd}+\beta_{tk}) - \log\Gamma(\beta_{tk})] + \sum_{d,k}[\log\Gamma(\sum_t \beta_{tk}) - \log\Gamma(\sum_t n_{tkd}+\beta_{tk})].$$

This is a separable function, since the first term depends only on α and the second only on β. Furthermore, the second term is a sum over topics, so each $\beta_{\cdot k}$ can be independently maximized. This gives a collection of $K+1$ equations to maximize:

$$\alpha'_{\cdot} = \text{argmax} \sum_{d,k}(\log\Gamma(n_{\cdot kd}+\alpha_k) - \log\Gamma(\alpha_k)) + \sum_{d}[\log\Gamma(\sum_k \alpha_k) - \log\Gamma(\sum_k n_{\cdot kd}+\alpha_k)]$$

$$\beta'_{\cdot k} = \text{argmax} \sum_{d,t}(\log\Gamma(n_{tkd}+\beta_{tk}) - \log\Gamma(\beta_{tk})) + \sum_{d}[\log\Gamma(\sum_t \beta_{tk}) - \log\Gamma(\sum_t n_{tkd}+\beta_{tk})].$$

Each equation above defines a vector, either $\alpha_{\cdot}$ or $\beta_{\cdot k}$. We use limited memory BFGS (Zhu et al., 1997) to perform the maximizations. For one iteration of EM with 20 topics on S&P500 data explained below, a careful Matlab implementation requires about 100 seconds on a 2.4GHz CPU with 6GB memory.

The implementation of DCMLDA allows both the α vector and β array to be non-uniform. For the DCMLDA model to be useful, β must be non-uniform, since it carries the information that ϕ carries in LDA. The vector α could be uniform in DCMLDA, but learning non-uniform values allows the model to give certain topics higher prior probability than others.

4. Experimental Design

Our experimental goal is to test whether the handling of burstiness in a DCMLDA model creates a better topic model than standard LDA. We compare DCMLDA models with LDA models, rather than with more complex topic models, for two reasons. First, DCMLDA and LDA are of comparable conceptual complexity. Second, and more important, they are competing models. DCMLDA is not in competition with more complex topic models, because these models can be modified to include DCM components.

Given a test set of documents not used for training, we estimate the held-out likelihood $p(w|\alpha,\beta)$ for LDA and DCMLDA models. The latter probability uses a vector $\alpha_{\cdot}$ and an array $\beta_{\cdot\cdot}$ learned as described above. The former probability uses $\alpha = \bar{\alpha}_{\cdot}$ and $\beta = \bar{\bar{\beta}}_{\cdot\cdot}$, the scalar means of the values learned by DCMLDA training. We also compare these two models to LDA using the values proposed by Griffiths and Steyvers (2004).

We compare LDA and DCMLDA as models for both text and non-text data. The textual dataset is a collection of papers from the 2002 and 2003 NIPS dataset compiled by Globerson et al. (2004) and organized by Elkan (2006). This dataset comprises 520955 words (6871 unique word types) in 390 documents. The second is a newly-compiled dataset of stock price fluctuations for the stocks that compose the S&P 500. This dataset contains 501 days of stock transactions between January 2007 and September 2008, with each document being a single day of trading. Each word is a concatenation of a stock symbol and a direction (+ or –), and each day contains one copy of a word for each (rounded) percentage point change between the opening and closing price of the stock. This dataset contains 469642 words in 501 documents. Both datasets are bursty, and approximately equally so, with $B = 2.63$ for NIPS and $B = 2.51$ for S&P500, where B is the burstiness measure from Church and Gale (1995), with $B = 1$ indicating no burstiness and higher values indicating more burstiness.

In analyzing the S&P500 data, the goal is to find groups of companies whose stock prices tend to move together. For example, a learned topic might hypothetically include the words IBM+, MSFT+, and AAPL–. This would indicate that IBM and Microsoft frequently rise together, while Apple tends to fall on the same days. Because different groups of stocks can move independently, each day can be a combination of a different set of topics.

5. Empirical Likelihood

Comparing the goodness-of-fit of topic models is a notoriously tricky endeavor. Ideally, we would calculate the incomplete likelihood $p(w|\alpha,\beta)$ for each model and compare those values. However, the incomplete likelihood is intractable for topic models. The complete likelihood $p(w,z|\alpha,\beta)$ is tractable, so previous work (Griffiths & Steyvers, 2004, e.g.) has calculated the harmonic mean of the complete likelihood from the topic assignments generated during Gibbs sampling. This approach is based on a true mathematical identity, but Newton and Raftery (1994) have argued that it is unreliable.

Another possibility is to measure classification accuracy, but that entwines the usefulness of the topics with the separability of the dataset. This is an important consideration because datasets do not always lend themselves to obvious classification schemes. Also, learned topics can be meaningful even if they are not well correlated with pre-assigned class labels.

We follow a third approach suggested by Li and McCallum (2006). This approach is to approximate the true held-out likelihood with so-called empirical likelihood (EL). To measure EL, we first train each model to obtain its parameter values α and β. These parameter values are then fed into the generative model, and a large set of pseudo documents is produced. Each of these documents has θ and ϕ distributions. (For DCMLDA the ϕ distribution of each document is different, while for LDA they are identical.) The pseudo documents are then used to train a tractable model. In the present case, we use a mixture of multinomials. Following Li and McCallum (2008), each multinomial model is inferred directly from the generated ϕ and θ distributions; individual words are not generated in the pseudo documents. The true likelihood of the test set is then estimated as its likelihood under the tractable model of the pseudo corpus. We report the arithmetic mean of log likelihoods of documents in the test set.

We investigate the stability of EL as a measure of goodness-of-fit by running it multiple times for the same DCMLDA model. Specifically, we train three separate 20-topic DCMLDA models on the S&P500 dataset, and run the EL method five times for each of these models. The mean absolute difference between EL values for the same model is 0.08%, with maximum 0.20%. Furthermore, the mean absolute difference between EL values for separately trained DCMLDA models is 0.11%, with maximum 0.29%, showing that likelihood values are stable over DCMLDA models with the same number of topics. The relationship between empirical likelihood and other measures of goodness-of-fit measures is unclear, but this stability suggests that EL is a sensible measure.

6. Results

An important, but informal, measure of the success of a topic model is the plausibility of the topics that it proposes. Since DCMLDA creates document-specific subtopics based on corpus-level topics, it is fair to ask if these corpus-level topics are as interpretable as LDA topics. Table 1 shows two topics from a 20-topic DCMLDA model trained on the S&P500 dataset. The words shown are the most likely based on the rank-order of the β_{tk} values over words t for a given topic k, in the same way that ϕ_{tk} indicates the most likely words for an LDA topic. The topics discovered by DCMLDA generally follow accepted stock classification systems. The 25 most likely stocks in the "computer related" topic are all in the Information Technology sector of the Global Industry Classification Standard (GICS), and 24 of the 25 most likely stocks in the "real estate" topic are in the Financials sector.

Table 1. Sample topics found by a 20-topic DCMLDA model trained on the S&P 500 dataset. The six most likely words for each topic are listed.

"Computer Related"		"Real Estate"	
Stock	Company	Stock	Company
NVDA+	Nvidia	SPG+	Simon Prop.
SNDK+	SanDisk	AIV+	Apt. Invest.
BRCM+	Broadcom	KIM+	Kimco Realty
JBL+	Jabil Circuit	AVB+	AvalonBay
KLAC+	KLA-Tencor	DDR+	Developers
NSM+	Nat'l Semicon.	EQR+	Equity Resid.

Table 2. Sample topics found by a 20-topic LDA model trained on the same S&P 500 dataset. The six most likely words for each topic are listed.

"Computer Related"		"Real Estate"	
Stock	Company	Stock	Company
NVDA+	Nvidia	LEN+	Lennar
SNDK+	SanDisk	CTX+	Centex
AMD+	AMD	PHM+	Pulte Homes
MU+	Micron	DHI+	D. R. Horton
BRCM+	Broadcom	KBH+	KB Home
CIEN+	Ciena	PLD+	ProLogis

The DCMLDA topics are similar to topics from a 20-topic LDA model trained on the same data, as shown in Table 2. Three of the top six companies in the computer topic are shared between the models. The LDA topic most similar to the DCMLDA "real estate" topic is also shown; all six top companies in the DCMLDA topic are among the top 15 of the LDA topic. Subjectively, the interpretability of the DCMLDA topics is comparable to the interpretability of the LDA topics. Looking closely suggests that the DCMLDA topics may be better. For example, all six top stocks for the DCMLDA "computer related" topic are suppliers to computer manufacturers, while Ciena in the matching LDA topic is not. In the LDA "real estate" topic the top five stocks are homebuilders but ProLogis is quite different. In contrast, all six stocks in the DCMLDA topic are corporate landlords.

As discussed in Section 5, we use empirical likelihood to compare the goodness-of-fit of the DCMLDA

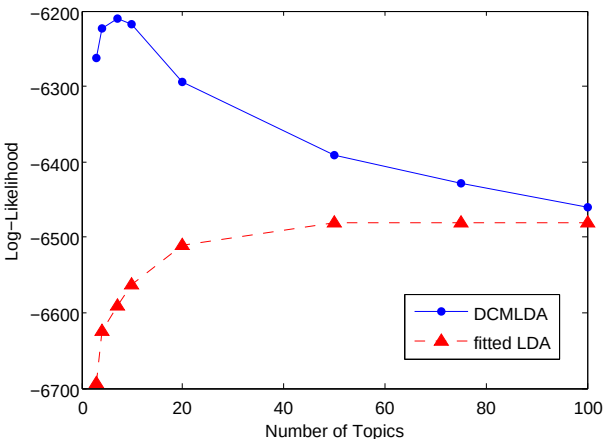


Figure 2. Mean per-document log-likelihood on the S&P500 dataset for DCMLDA and fitted LDA models. The heuristic model is omitted here because its likelihood is too low. The maximum standard error is 11.2.

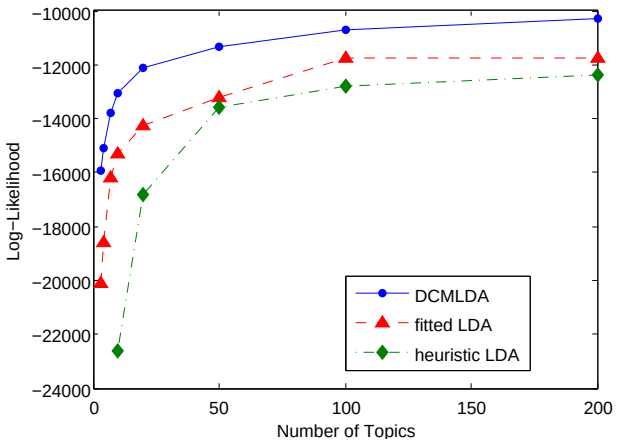


Figure 3. Mean per-document log-likelihood on the NIPS dataset for DCMLDA and LDA models. The maximum standard error is 183.6.

and LDA models on the NIPS and S&P500 datasets. We perform five 5-fold cross-validation trials for each number of topics and each dataset. We first train a DCMLDA model, then create two LDA models. One ("fitted LDA") uses the mean values of the DCMLDA hyperparameters. The other ("heuristic LDA") uses the uniform hyperparameter values suggested by Griffiths and Steyvers (2004). For both datasets, DCMLDA is better than fitted LDA, which in turn is better than heuristic LDA.

Figure 2 shows performance on the S&P500 dataset. The highest likelihood comes from DCMLDA with seven topics, where DCMLDA has a major advantage over the fitted LDA model. This supports the idea that a DCMLDA model with few topics is comparable to an LDA model with many topics. This may also indicate that the a natural set of topics for this dataset has cardinality about seven.

Above 100 topics, the likelihood of the fitted LDA model remains approximately constant, while that of DCMLDA continues dropping, ending up lower than that of LDA. This is likely a result of data sparsity preventing the estimation of good β values. As there are only 1000 unique symbols in the dataset, poor behavior with more than 100 topics is not a major source of concern. The likelihoods for heuristic LDA model are not shown in Figure 2 because they are much lower than those of the other models, especially when the number of topics is low. For 100 topics, heuristic LDA has mean log-likelihood -7383, which approaches that of the other two models, but for three topics, its mean log-likelihood is -34130.

Figure 3 shows performance on the NIPS dataset. For this dataset, the DCMLDA model does not exhibit the few-topics bump seen in the S&P500 dataset. DCMLDA outperforms the fitted LDA model at every tested number of topics. For the NIPS dataset, LDA never surpasses DCMLDA as the number of topics grows, presumably because the larger number of unique words (6871) in this corpus keeps data sparsity from becoming a major issue. LDA with the heuristic hyperparameter values is not as bad on the NIPS dataset as on the S&P500 dataset, almost catching up with the fitted LDA model at 50 topics. This confirms that the suggestions of Griffiths and Steyvers (2004) are reasonable for textual data. However, the fitted LDA model retains a substantial advantage, especially when the number of topics is small.

7. Discussion

While the choice of α and β in a topic model is sometimes viewed as a formality, and heuristic values are used without much consideration, we find that heuristic values can lead to much worse likelihood than fitted values, especially when the number of topics is small. Thus learning α and β can be beneficial, and optimized values can be significantly different from previously suggested heuristic values. In addition, we see that accounting for burstiness improves held-out likelihood for both text and non-text data. To be completely confident that the EL improvement is due to modeling burstiness, DCMLDA should be compared also to a version of LDA with a single optimized non-uniform β parameter.

Recent years have seen a profusion of topic model variants, such as the correlated topic model (Blei & Lafferty, 2005) and the Pachinko allocation model (Li & McCallum, 2006). These newer models outperform LDA on many tasks, so comparing the performance of DCMLDA only to that of LDA may seem inappropriate. However, DCMLDA is not in competition with the more complex topic models, but rather with LDA. The more complex topic models share an LDA core, in that they use multinomials to represent topics. These multinomials can be replaced by DCMs to improve, potentially, the performance of these models. Thus the DCMLDA idea and complex topic models are complementary.

Acknowledgments. The first author was supported in part by NIH Training Grant T32-DC000041. We wish to thank the UCSD Computational Psycholinguistics Lab for insights and advice.

References

Airoldi, E. M., Fienberg, S. E., & Xing, E. P. (2007). Mixed membership analysis of genome-wide expression data. Arxiv preprint arXiv:0711.2520.

Blei, D., & Lafferty, J. (2005). Correlated topic models. *Advances in Neural Information Processing Systems 18* (pp. 147–154).

Blei, D., Ng, A., & Jordan, M. (2001). Latent Dirichlet allocation. *Advances in Neural Information Processing Systems 14* (pp. 601–608).

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *J. of Machine Learning Research*, *3*, 993–1022.

Celeux, G., Chaveau, D., & Diebolt, J. (1996). Stochastic versions of the EM algorithm: An experimental study in the mixture case. *J. of Statistical Computation and Simulation*, *55*, 287–314.

Church, K., & Gale, W. A. (1995). Poisson mixtures. *Natural Language Engineering*, *1*, 163–190.

Elkan, C. (2006). Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. *Proceedings of the 23rd International Conference on Machine Learning* (pp. 289–296).

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 524–531).

Globerson, A., Chechik, G., Pereira, F., & Tishby, N. (2004). Euclidean embedding of co-occurrence data. *Advances in Neural Information Processing Systems 17* (pp. 497–504).

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *104*, 5228–5235.

Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2004). Integrating topics and syntax. *Advances in Neural Information Processing Systems 17*.

Heinrich, G. (2005). Parameter estimation for text analysis. Available at `http://www.arbylon.net/publications/text-est.pdf`.

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. *Proceedings of the 23rd International Conference on Machine Learning* (pp. 577–584).

Li, W., & McCallum, A. (2008). Pachinko allocation: Scalable mixture models of topic correlations. *J. of Machine Learning Research*. Submitted.

Madsen, R., Kauchak, D., & Elkan, C. (2005). Modeling word burstiness using the Dirichlet distribution. *Proceedings of the 22nd International Conference on Machine Learning* (pp. 545–552).

Newton, M., & Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society B*, *56*, 3–48.

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive Bayes text classifiers. *Proceedings of 20th International Conference on Machine Learning* (pp. 616–623).

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, *23*, 550–560.