Selecting the Number of Topics in a Latent Dirichlet Allocation Model

Dale Bowman*

Abstract

Topic modeling is a useful tool for examining latent structures in a corpus of documents. Latent Dirichlet Allocation (LDA) is a popular topic modeling method that assumes a Bayesian generative model for collections of exchangeable binary observations such as the presence or absence of words within a document. The degree to which an LDA model is useful for modeling a corpus depends, in part, on the number of topics selected. Too few topics can result in an LDA model that does not provide sufficient separation of topics and too many topics can result in a model that is overly complex and difficult to interpret. Several ad hoc, heuristic methods for selecting the proper number of topics have been proposed. These typically require that the LDA model be fit over a varying number of topics and the performance of the resulting model be measured by some criteria such as perplexity, rate of perplexity change, and goodness of fit statistics.

Key Words: Exchangeability, topic models, LDA, perplexity, goodness of fit

1. Introduction

Topic models are tools used to uncover latent topics in collections of discrete objects. These tools were developed to deal primarily with text processing. Three major topic modeling methods are: Latent Dirichlet Allocation (LDA) (Blei et al., 2003), probabilistic Latent Semantic Indexing (pLSI) (Hoffman, 1999), and Latent Semantic Indexing (LSI) (Deerwester et al., 1990). The LDA model has become the most popular of the three due, in part, to it's unsupervised nature and to the ease with which LDA may be extended to any collection of discrete data, including text analytics (eg. Griffiths and Stevens, 2004, Blei et al., 2003, Blei and Lafferty, 2007), image retrieval (Blei and Jordon, 2003), social network analysis (Airoldi et al., 2008), and bioinformatics (eg. Rogers et al., 2005, Shivashankar et al., 2011, Zhao et al., 2014, and Coelho et al., 2010).

Following conventions of language processing, the discrete objects are termed documents, and their collection is a corpus. Each document consists of a list of words assumed to be exchangeable within each document. It is assumed that documents are independent of each other and are generated by a Bayesian process. The LDA model assumes that words are generated by first selecting a topic and then selecting a word from the distribution of words specific to the topic chosen. The probabilities for topic choice and word choice given topic follow multinomial distributions assumed to be conditionally independent. The parameters of the multinomial distributions are chosen from prior Dirichlet distributions. LDA is typically used to cluster words within documents and to reduce the dimension of the documents from a large vocabulary space to a smaller dimensioned topic space.

The ability of LDA to successfully cluster document collections into meaningful themes has been demonstrated for document collections, where there exists some "ground truth" to compare with the LDA classification (Zhao et al., 2014). In the case where a "ground truth" is known to exist, the number of themes or topics may be known *a priori*. In the absence of the knowledge of the "true" number of topics, the best number of topics to use in the LDA model is not known. Different numbers of topics are likely to induce very different structures in the collection of documents. Too few topics could result in a classifier that is

*Department of Mathematical Sciences, The University of Memphis, Memphis TN 38152

too coarse to be useful and too many topics could result in overfitting - resulting in a model that is too complex for ease of interpretation.

Most procedures for determining the optimum number of topics for a corpus rely on selecting a candidate set of topic numbers, fitting LDA models with each topic number from the candidate set, and using some criteria to determine which number of topics provides the best fit. Most of these procedures are heuristic in nature. One measure of the LDA fit that is used to determine the best number of topics for a corpus selects the number of topics for which the model is least perplexed. The perplexity was suggested by Blei et al., (2003) as a measure of model fit. Other methods that have been proposed include the use of rate of perplexity change (Zhao et al., 2015), the empirical method of Arun et al., (2010), the use of a cosine similarity measure (Grant et al., (2013), using a goodness of fit measure (Bowman et al., 2016) and using pointwise mutual information (PMI) (Song et al., 2009). In contrast to these ad hoc methods for selecting the optimum number of topics, Teh et al., (2006) propose a hierarchical LDA model that uses a Dirichlet process prior to estimate the number of topics within the estimation procedure. In this paper, heuristic methods are discussed and compared. The remainder of this paper is organized as follows. An introduction to LDA inference and estimation is given in Section 2. Section 3 discusses several proposed methods for selecting an optimum number of topics and discusses a new method based on a goodness of fit test. Two methods are compared on real corpora in Section 4 and a discussion of results is given in Section 5.

2. Latent Dirichlet Allocation

The LDA model, a generative hierarchical Bayesian model, was introduced by Blei et al., (2003) for collections of binary data. LDA assumes that words within a document are exchangeable, (the "bag of words" assumption) and that the topic and word distributions are conditionally independent. The described model is based on the assumption that a document (a collection of binary data) is generated as follows from a set of K topics and a vocabulary of V words. Each word in the document is generated by

1. Selecting a $K \times 1$ vector, $\boldsymbol{\theta}$, from a Dirichlet distribution with parameters $\boldsymbol{\alpha}$.
2. Using the parameters $\boldsymbol{\theta}$, a single topic, say topic i, is chosen from a Multinomial distribution.
3. For the selected topic a $V \times 1$ vector, $\boldsymbol{\phi}_i$ is selected from another Dirichlet distribution with parameters $\boldsymbol{\beta}_i$ dependent upon the topic selected in the previous step.
4. A word is chosen from a multinomial distribution with parameters $\boldsymbol{\phi}_i$ conditioned on the topic chosen.

This process is repeated for each word in the document. Figure 1 shows a graphical representation of the generative LDA process. It can be seen from Figure 1 that the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are corpus level parameters and chosen once for the corpus, the $\boldsymbol{\theta}$ parameters are sampled for each document and the variables $\boldsymbol{Z}_n$ and $\boldsymbol{p}$ are sampled for each word in each document.

Given the generative process, the LDA model for a single document may be formulated as follows. Define a vector of binary variables $\boldsymbol{Z}_n = (Z_{n1}, Z_{n2}, \ldots, Z_{nK})$ as the vector indicating which topic the nth word is chosen from. Here

$$Z_{ni} = \begin{cases} 1 & \text{if the nth word is from topic i} \\ 0 & \text{otherwise.} \end{cases}$$

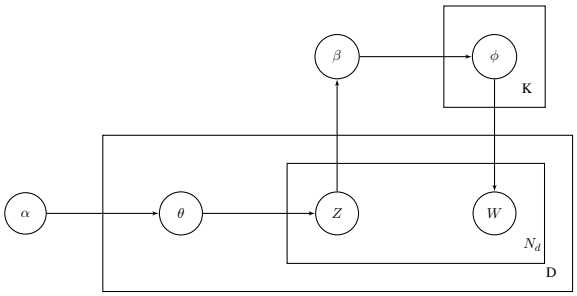

Figure 1: LDA Generative Model

The likelihood of $\boldsymbol{Z}_n$ is a multinomial given parameter $\boldsymbol{\theta}$ and the joint likelihood of $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N|\boldsymbol{\theta})$ is

$$L(\boldsymbol{Z}|\boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{i=1}^{K} \theta_i^{Z_{ni}},$$

where N is the number of words in the document. Here θ_i is the probability of the ith topic within the document for $i = 1, \ldots, K$. Similarly define $\boldsymbol{w}_n = (w_{n1}, \ldots, w_{nV})$ as the vector indicating which vocabulary word was chosen for the nth word. Assuming conditional independence given $\boldsymbol{Z}$ and $\boldsymbol{\phi}$, the likelihood of $(\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_N)$ is then

$$L(\boldsymbol{w}|\boldsymbol{Z}, \phi) = \prod_{n=1}^{N} \prod_{i=1}^{K} \prod_{j=1}^{V} \phi_{ij}^{Z_{ni} W_{nj}},$$

where ϕ_{ij} is the word probability distribution over the vocabulary specific to the ith topic. The prior distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are Dirichlet distributions with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively, resulting in a joint posterior distribution of the latent variable $\boldsymbol{Z}$ and parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$,

$$\pi(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \prod_{n=1}^{N} \prod_{i=1}^{K} \left(\theta_i^{(Z_{ni} + \alpha_i - 1)} \prod_{j=1}^{V} \phi_{ij}^{(Z_{ni} w_{nj} + \beta_{ij} - 1)} \right).$$

The marginal posterior distribution of words in a document is intractable and estimation schemes such as variational inference and Markov chain Monte Carlo (MCMC) are used to estimate parameters and posterior distributions.

3. Methods for Selecting the Optimum Number of Topics

When the LDA model is assumed for a corpus, both variational inference and MCMC methods require the number of topics to be specified in advance. The quality of the LDA model fit is, in large part, determined by choosing an appropriate number of topics. Most of the methods proposed in the literature evaluate model fit over a range of candidate values for K, the number of topics. The "optimum" number of topics is then chosen as the number that provides the best model as specified by some criteria. Such approaches to obtaining a good number of topics for a particular corpus rely on a time consuming trial and error approach. In contrast, the methods derived by Teh et al., (2006) estimate the number of topics from the corpus using hierarchical LDA that assumes a Dirichlet process prior to estimate the number of topics. The hierarchical LDA model has not been found useful in practice and is not computationally efficient (Wallach et al., 2009). Researchers tend to use a reasonable guess for the number of topics to use or rely on ad hoc procedures as discussed next.

3.1 Perplexity

Perplexity is a measure of model fit often used in language processing. Blei et al., (2003) suggest using perplexity to assess LDA model fit. If a certain proportion of documents are randomly held out from the corpus in order to test the model fit on the training documents, then the perplexity of the test set of documents provides a measure of how well the LDA model was trained on the corpus. The lower the perplexity score the better the model fit. The perplexity is defined on a test set, D_{test} of size M as

$$perplexity(D_{test}) = exp\left(-\frac{\sum_{d=1}^{M}\log(p(w_d))}{\sum_{d=1}^{M}N_d}\right). \quad (1)$$

To use the perplexity to select the number of topics to use for a particular corpus, the perplexity would be computed on test sets from trained LDA models fit over a set of candidate topics. The number of topics for which the trained model is least perplexed is chosen as the optimal number of topics.

The use of perplexity to obtain a good number of topics for a particular corpus may provide meaningful results. However, the perplexity tends to be unstable for some corpora with results varying for the same corpus with different starting seeds (Zhao et al., 2015). In addition, Figure 1 shows a typical plot of perplexity by number of topics. As seen in the figure, the perplexity is large for a small number of topics (left of the red line) and falls off rapidly. The perplexity then tends to increase after some number of topics (right of the green line) perhaps indicating overfitting of the LDA model. The perplexity in the area between is low and relatively the same for all K in this region. This indicates that any number of topics in this region (between red and green lines) would provide a model with adequate fit. Leading to the question of whether to choose the number of topics which yields the absolute minimum perplexity or in the name of parsimony should the smallest number of topics within this valley be chosen? Clearly the use of perplexity requires some additional guidelines when using it to determine an optimum number of topics.

3.2 Rate of Perplexity Change

In order to avoid some of the problems that may occur when using perplexity to determine the number of topics, Zhao et al., (2015) propose the use of rate of perplexity change, RPC. For a candidate set of topics in increasing order $(k_1, \ldots, k_t)$ and perplexities $(P_1, \ldots, P_t)$ from corresponding LDA models, the RPC is defined for topic k_i as

$$RPC(i) = \left|\frac{P_i - P_{i-1}}{k_i - k_{i-1}}\right|. \quad (2)$$

The first i that satisfies $RPC(i) < RPC(i+1)$ is chosen as the optimal number of topics.

The stability of the RPC based method was compared to that of the perplexity based method by Zhao et al., (2015) using Shannon entropy. The RPC method was consistently found to be more stable.

3.3 Topic Coherence

A good LDA model should result in topics with high coherence. In Newman et al., (2010) methods for evaluating topic coherence were compared and evaluated using large corpora from Wikipedia, Google n-gram datasets and WordNet. They found that pointwise mutual information (PMI) provided the best method for measuring topic coherence when compared to human judgement. The PMI of two words, A and B is defined to be

$$PMI(A, B) = log\left(\frac{P(AB)}{P(A)P(B)}\right), \quad (3)$$

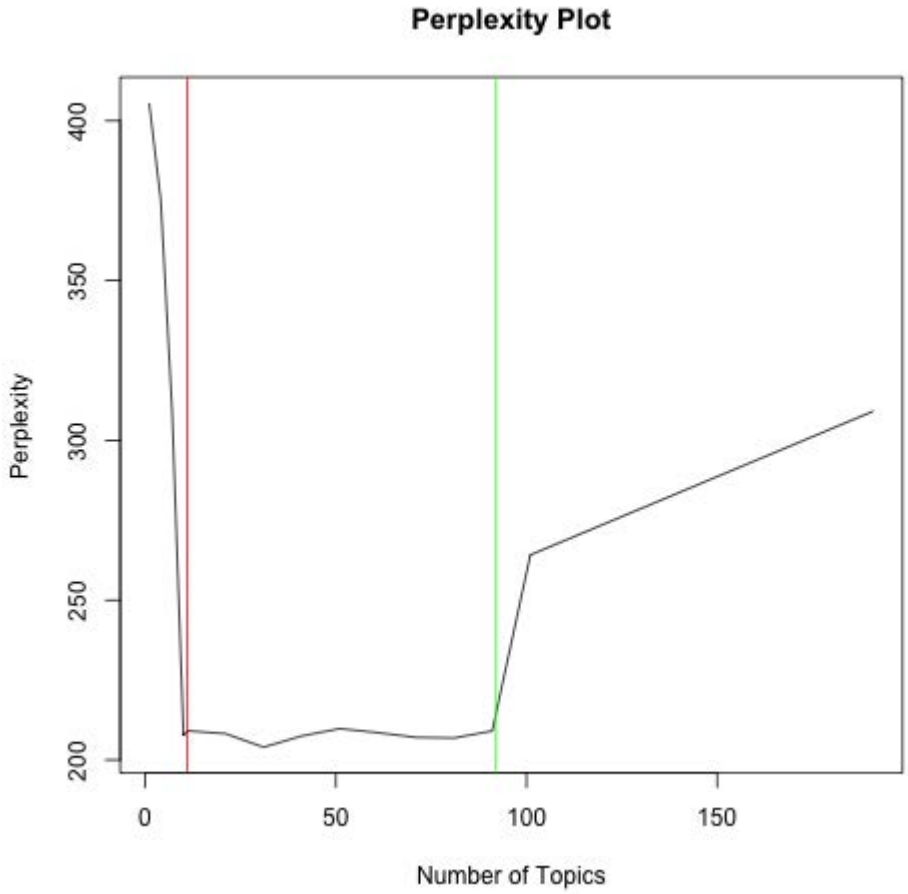

Figure 2: A Typical Plot of Perplexity by Number of Topics

where $P(x)$ is the probability of event x occurring and $P(xy)$ is the probability of both x and y occurring. The number of topics from a candidate set that results in the highest measure of PMI for a topic model is considered the best number of topics for that corpus. In order to obtain the PMI values for trained LDA models, the top m number of words within each topic are considered. The top words are found using the estimated word given topic distribution from the fitted LDA model, either using m as a fixed number (Niraula et al., 2013) or as the number of words with probability exceeding a cutoff level. For all possible pairs of the top m words per topic the PMI is calculated using a large external reference set. In Niraula et al., (2013) for example the external reference set was based on word frequencies derived from 4,134,837 Wikipedia articles. The total PMI is found for the corpus and the optimum number of topics is considered the one with largest generated PMI value.

3.4 Goodness of Fit

A proposed method for determining a preferred number of topics is based on a goodness of fit test derived by Bowman et al., (2016). A measure of the goodness of fit of topic assignment tests the assumption that topics are randomly assigned to documents. It is presumed that an LDA model provides good fit to a corpus if topics are not randomly assigned to a predominant number of documents in the corpus. This idea can be used to

select the number of topics by finding the proportion of documents in the corpus that fail to reject a random assignment hypothesis for each of a set of candidate topic numbers. The "best" number of the topics is the number with fewest random topic assignments.

Let $\boldsymbol{\theta}^d = (\theta_1^d, \theta_2^d, \ldots, \theta_K^d)'$ be the vector of topic probabilities in the dth document of the corpus for a model with K topics. Thus θ_i^d may be interpreted as the probability that topic i is included in document d. The goodness of fit null hypothesis in document d is $H_0 : \theta_i^d = \theta_j^d = 1/K$ for $i, j = 1, \ldots, K$. Let Z_i^d be the number of words in document d that are from topic i. Then $(Z_1^d, \ldots, Z_K^d)$ follows a multinomial distribution with $(N_d, \boldsymbol{\theta}^d)$ where N_d is the number of words in document d. Under the null hypothesis, the likelihood is maximized when $\hat{\theta}_i^d = 1/K$ for $i = 1, \ldots, K$.The likelihood ratio function is then given by

$$\lambda = \prod_{d=1}^{D} \prod_{i=1}^{K} \left(\frac{N_d}{K Z_i^d} \right)^{Z_i^d}. \tag{4}$$

The log likelihood ratio test statistic for the dth document is then

$$T_d = 2 \sum_{i=1}^{K} Z_i^d \log \left(\frac{Z_i^d K}{N_d} \right). \tag{5}$$

Under the null hypothesis T_d can be compared to a chi-square distribution with $K - 1$ degrees of freedom. For each $d = 1, \ldots, D$ document a GOF test of random assignment is performed and the proportion of documents which failed to reject random assignment is obtained. The candidate topic number with smallest proportion of documents with random assignment is considered the best number of topics to use with LDA for that corpus.

4. Examples

The goodness of fit approach and the rate of perplexity change are compared using two different data sets. The first data set was retrieved from the publicly available SIDER2 database (http://sideeffects.embl.de). The data set is discussed in Kuhn et al., (2010). The data base contains 996 drugs, each of which are considered documents in this study. Associated with each drug are self reported side effects from a vocabulary of 3,034 possible adverse events after some pre-processing. The side effects reported for a particular drug are the "words" in the document. The dataset can be envisioned as a 996×3034 matrix of 1's and 0's. A one in position a_{ij} in this matrix indicates that for the ith drug the jth side effect has been associated with that drug.

Topic models with varying number of topics were fit to this data set using the LDA algorithm in Mallet (McCallum, 2002). The rate of perplexity change method of Zhao et al., (2015) selected 50 topics as the optimum number for this data set. A graph of rate of perplexity change for the SIDER2 corpus is given in Zhao et al., (2015). Table 1 shows the results from the goodness of fit method for determining the appropriate number of topics. The table shows, for a range of number of topics, the proportion of the 996 documents for which the null hypothesis of random topic assignment was not rejected at the $\alpha = .05$ level. The proportion of randomly assigned topics is small when the number of topics, K is ten and then increases at $K = 20$, falling off to a low of 0.006 when the number of topics is 40. The proportion then increases afterwards, with 30% of the documents subject to random assignment when $K = 100$. The goodness of fit method would suggest that $K = 30$ is the best number of topics to use for this data set. Although the two methods propose different number of topics, it is seen in Zhao et al., (2015) that the performances of the topic model when $K = 50$ as chosen by the RPC and the topic model when $K = 40$ as chosen by

Table 1: Proportion of Documents with Random Assignment - SIDER2

K	10	20	30	40	50	100
Proportion	0.0291	0.1757	0.0833	0.0060	0.1546	0.3042

the goodness of fit method are similar. The goodness of fit method is based on a statistical procedure whereas the RPC is a heuristic approach. Table 1 shows that if $K = 50$ topics are used, 15% of the documents are likely to have random topic assignment, suggesting over-fitting. The perplexity alone was not found to be stable enough to use to select an optimum number of topics for this corpus.

The second data set was created by Zhao et al., (2015). Abstracts of papers published in the IEEE Transactions on Computational Biology and Bioinformatics (TCBB) from the PubMed database were used as documents in this corpus. The number of abstracts was 885 from papers published between 2004 and 2013. After pre-processing and removal of stopwords (common English adverbs, conjunctions, pronouns and prepositions), there were 5004 words in the vocabulary. Table 2 shows the proportion of the 885 documents for which the goodness of fit test for random assignment failed to reject at $\alpha = 0.05$ level of significance. The table shows that there were no documents with randomly assigned topics at $K = 10$ and the proportion of documents with randomly assigned topics increases gently afterwards. For any of $K = 20, 30, 40$ and 50 topics the proportion of random assignment seems acceptable, increasing to an unacceptable level of 0.1571 when the number of topics is 100. Zhao et al., (2015), using the rate of perplexity change, found the best number of topics for this data set to be 40. Since the pattern of proportions of documents with random assignment does not suggest a "best" number of topics using the goodness of fit method, a second goodness of fit is proposed to refine the choice. To this end, the last line in Table 2 shows the proportion of topics for which a goodness of fit test of the null hypothesis that words are randomly assigned to topic fails to reject at $\alpha = 0.05$. When $K = 10$ and $K = 20$, none of the topics are judged to have words randomly assigned. However, when the number of topics is 30, more than a third (11 out of 30 = 36.7%) of the topics have word distributions that are not different from a random (uniform) word distribution. The test for goodness of fit of words per topic is derived in Bowman et al., (2016). Using both goodness of fit measures for this data leads to the conclusion that $K = 20$ topics would be the best to use.

Word clouds are used by Zhao et al., (2015) to represent the LDA-derived topic-words distributions for this corpus. Word clouds with differing number of topics can be compared for cohesiveness within topics and for distinguishability between topics. Using the word clouds, Zhao et al., (2015) argue that the optimum number of topics selected by the RPC method, 40, gives topics with unique and distinguishable themes. They argue that for the optimum number of topics selected by the goodness of fit measure, 20, some topics were easily identified by theme and were distinguishable from others, although some topics tended to combine themes that were distinguished by the 40-topic model. The measures of the word distribution within topics using word clouds is a subjective method. Balance should be weighted between topic cohesion and distinction and parsimony and goodness of fit when selecting an optimum number of topics.

Table 2: Proportions of Random Assignment - TCBB

K	10	20	30	40	50	100
Document Proportions	0.0000	0.0023	0.0056	0.0113	0.0294	0.1571
Topic Proportions	0.0000	0.0000	0.3670	0.4750	0.5600	0.8300

5. Discussion

LDA topic modeling can offer a useful and effective tool for exploring latent structures in large collections of documents. The degree to which an LDA model can be effective is related to the choice of an appropriate number of latent topics. In this paper, several proposed methods for estimating a "best" number of topics for a particular corpus have been discussed and compared. A new method for selecting the optimum number of topics based on a goodness of fit test is proposed and discussed. In examples to corpora in which the perplexity itself was too unstable to be useful in obtaining an optimum number of topics, the goodness of fit method and a method based on rate of perplexity change showed comparable results. The advantage of the goodness of fit measure is that it is an inference-based diagnostic, where the RPC method is a heuristic approach without inferential support. Neither procedure has computational advantages over the other, nor can either method claim optimality in the usual sense. The proposed pointwise mutual information method to find an optimum number of topics is based on the concept of selecting as optimum the number of topics with the most cohesion within topics. The limitation of this method is that an external reference set is needed to compute the probabilities of pairs of words occurring together and of single word occurrences. The external reference set needs to be large enough to give meaningful estimates of the required probabilities and may be discipline dependent.

Further work is in progress on improved methods of estimating pairwise mutual information and of obtaining maximum likelihood estimators for the number of topics for a corpus.

6. Acknowledgements

This work was supported in part by appointments to the Faculty Research Participation program at the National Center for Toxicological Research administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration.

REFERENCES

Airoldi, E.M., Blei, D.M., Feinberg, S.E., and Xing, E.P. (2008), "Mixed Membership Stochastic Block Models", *Journal of Machine Learning Research*, 9, 1981–2014.

Arun, R., Suresh, V., Madhaven, C.V., and Murthy, M.N. (2010), "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations", *Advances in Knowledge Discovery and Data Mining*, 391–402.

Blei, D.M. and Jordan, M.I. (2003), "Modeling Annotated Data", *The Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 127–134.

Blei, D.M. and Lafferty, J.D. (2007), "A Correlated Topic Model of Science", *The Annals of Applied Statistics*, 1, 17–35.

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003), "Latent Dirichlet Allocation", *The Journal of Machine Learning Research*, 3, 993–1022.

Bowman, D, Chen, J.J., and George, E.O. (2016), "Goodness of Fit Measures for Topic Modeling", *submitted*.

Coelho, L.P., Peng, T., and Murphy, R.F. (2010), "Quantifying the Distribution of Probes between Subcellular Locations Using Unsupervised Pattern Unmixing", *Bioinformatics*, 26, 7–12.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990), "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science*, 41, 391–407.

Grant, S., Cordy, J.R., and Skillicorn D.B. (2013), "Using heuristics to estimate an appropriate number of latent topics in source code analysis", *Science of Computer Programming*, 78, 1663–1678.

Griffiths T.L. and Stevens, M. (2004), "Finding Scientific Topics", *Proceedings of the National Academy of Sciences of the United States of America*, 101, 5228–5235.

Hoffmann, T. (1999), "Probabilistic Latent Semantic Indexing", *in Proceedings of SIGIR'99*, 50-57.

Kuhn, M., Campillos M., Letunic I., Jensen L.J., and Bock, P. (2010), "A side effect resource to capture phenotypic effects of drugs", *Molecular Systems Biology*, 6:343.

McCallun, A.K. (2002), "Mallett: A Machine Learning for Language Toolkit", *http://malletcsumassedu*.

Newman, D., Lau, J.H., Grieser, K., and Baldwin, T. (2010), " Automatic evaluation of topic coherence", *in HLT-NACL*, 100–108.

Niraula, N., Banjade, R., Stefanescu, D., and Rus, V. (2013), " Experiments with Semantic Similarity Measures based on LDA and LSA", *Proceedings of SLSP 2013*, 188–199

Rogers, S., Girolami, M., Campbell, C. and Breitling, R. (2005), "The Latent Process Decomposition of cDNA Microarray Data Sets", *IEEE ACM Transactions on Computational Biology and Bioinformatics*, 2, 143-156.

Shivashankar, S., Srivathsan, S., Ravindarn, B. and Tendulkar, A.V. (2011), "Multi-View Methods for Protein Structure Comparison Using Latent Dirichlet Allocation", *Bioinformatics*, 27, 161–168.

Song, Y., Pan, S., Liu, S., Zhou, M.X., and Qian, W. (2009), "Topic and Keyword Re-ranking for LDA-based Topic Modeling", *CIKM '09 Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 1757-1760.

Teh, Y.W., Jordan, M.I., Beal, M.J., and Blei, D.M. (2006), "Hierarchical Dirichlet Processes", *Journal of the American Statistical Association*, 101, 1566–1581.

Wallach, H., Mimno, D., and McCallum, A. (2009), "Rethinking LDA: Why Priors Matter?", *Advances in Neural Information Processing Systems*, 22, 1973–1981.

Zhao, W., Chen, J.J., Liu, Z., Ge, W., Ding, Y. and Zou, W. (2015), "A Heuristic Approach to Determine the Optimal Number of Topics in Topic Modeling", *BMC Bioinformatics*, 16(Suppl 13):S8.

Zhao, W., Zou, W., and Chen, J.J. (2014), "Topic Modeling for Cluster Analysis of Large Biological and Medical Datasets", *BMC Bioinformatics*, 15 Suppl. 11:S11.