BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–5

Goodness of Fit Measures for Topic Modeling

Bowman D [1,*], Chen JJ [2] and George EO [1*]

[1]Department of Mathematical Sciences, University of Memphis, Memphis TN 38152
[2]Department of Biostatistica and Bioinformatics, National Center for Toxicological Research, Jefferson, AT

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Topic modeling has wide applications to exchangeable discrete data composed of individual *documents*. Interest is in summarizing large datasets with a specified set of topics or in clustering the data into meaningful groups. Latent Dirichlet Allocation (LDA) topic modeling is a Bayesian generative model used for both purposes. The degree to which a trained model fits a dataset has typically been measured using the perplexity score computed on a held out set of documents. Several goodness of fit measures are proposed in this paper for testing the degree to which a trained LDA model fits documents in a corpus.

Results: The goodness of fit measures are illustrated on two different types of datasets involving abstracts of papers published in the IEEE Transations on Computation Biology and Bioinformatics (TCBB) from the PubMed database and a dataset of pharmacological side effects, publicly available at http://sideeffects.embl.de.

Availability and Implementation: Publicly available software MALLET at http://malletcsumassedu was used to perform the LDA analysis on the datasets. The goodness of fit tests were performed on outputs from this software in R.

Contact: ddbowman@memphis.edu

1 INTRODUCTION

Topic modeling is a tool that has been used for the statistical analysis of collections of documents. The goal in topic modeling is to summarize members of the collection by the estimation of latent topics within the documents. The latent Dirichlet allocation (LDA) topic model of Blei *et al.*, 2003 describes a Bayesian process by which documents are generated from multinomial distributions with Dirichlet prior distributions. The words are assumed exchangeable within topics and the topics are exchangeable within documents. Conditional independence is assumed by Blei *et al.*, 2003 in order to specify the model but this assumption is somewhat relaxed in later work by Blei and Lafferty, 2007.

Topic modeling has applications in a wide variety of areas involving discrete observations. Fields such as text mining (eg, Hofmann, 2001, Griffiths and Steyvers, 2004, Blei *et al.*, 2003, and Blei and Lafferty, 2007), image retrieval (Blei and Jordan, 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) have all used topic modeling for analyzing big data sets for feature extraction and feature selection.

The most commonly used means of assessing the fit of the LDA topic model is the calculation of the perplexity, a convention from language modeling. The perplexity is computed on a held out data set after the topic model has been trained on another set. It is most useful in comparing the fit of different models, for example, in determining which number of specified topics results in the best fit. In this paper several alternate measures are proposed for assessing the fit of the trained topic model, including goodness of fit tests and information criteria based on posterior distributions.

2 APPROACH

Consider a corpus consisting of D documents with the dth document consisting of N_d words selected from a vocabulary of size V. There are K latent topics distributed across each document with probabilities $\boldsymbol{\theta}^{(d)}$. Given that a word belongs to a specific topic, the probability of each vocabulary word appearing in the document follows a topic specific distribution with parameters $\boldsymbol{\phi}$. The LDA model assumes conditional independence of topic given $\boldsymbol{\theta}$ and of word given topic and $\boldsymbol{\phi}$. The described model is based on the assumption that a single document is generated as:

1. A $K \times 1$ vector, $\boldsymbol{\theta}$, is chosen from $Dir(\boldsymbol{\alpha})$.
2. The nth word is chosen from the following process for $n = 1, \ldots, N$ where N is the number of words in the document:

   a. Choose a topic Z_n from $Mult(1, \boldsymbol{\theta})$.

   b. Choose a $V \times 1$ vector, $\boldsymbol{\phi}_{z_{ni}}$ from $Dir(\boldsymbol{\beta}_{z_{ni}})$. Where this vector is dependent on the Z_{ni} chosen in the previous step.

   c. Choose a word w_n from $p(w_n|Z_n, \boldsymbol{\phi}_{z_{ni}})$ which is a $mult(1, \phi_{1z_{ni}}, \ldots, \phi_{Vz_{ni}})$.

In the generative process $\phi_{ij} = Prob(w_{nj} = 1|z_{ni} = 1)$ where $w_{nj} = 1$ if the nth word in the document is the jth word in the vocabulary and this value is observed. $z_{ni} = 1$ if the nth word in the document is in topic i and is unobserved. Note

$$\sum_{i=1}^{K} z_{ni} = \sum_{j=1}^{V} w_{nj} = 1.$$

[*]to whom correspondence should be addressed

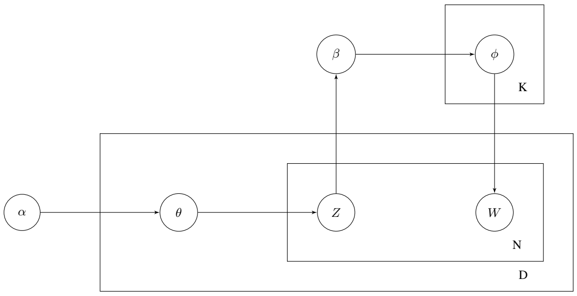

Fig. 1. LDA Generative Model

This process is summarized in the graph in Figure 1 where it can be seen that the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are corpus level parameters sampled (or assigned) only once when generating a set of documents. The parameters $\boldsymbol{\theta}$ are document specific parameters, sampled once for each document. Variables $\boldsymbol{Z}$ and $\boldsymbol{w}$ are sampled once for each word in each document while $\boldsymbol{\phi}$ are topic level variables sampled for each topic based on the sampled value of $\boldsymbol{Z}$.

Given this generative process, the LDA model can be written as follows. In a single document, $\boldsymbol{Z}_n = (Z_{n1}, \ldots, Z_{nK})$ is the vector indicating which topic the nth word belongs to. As assumed in the generative model $\boldsymbol{Z}_n \sim Mult(1, \theta_1, \ldots, \theta_K)$ for $n = 1, \ldots, N$. Hence

$$f(\boldsymbol{Z}_n|\boldsymbol{\theta}) = \prod_{i=1}^{K} \theta_i^{z_{ni}}$$

for $\sum_{i=1}^{K} z_{ni} = 1$ and $\sum_{i=1}^{K} \theta_i = 1$ and

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

For each document θ_i can be interpreted as the probability topic i is in the document.

Let $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)$. Then the likelihood of $\boldsymbol{Z}$ given $\boldsymbol{\theta}$ is

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

due to the assumption of conditional independence of topic given $\boldsymbol{\theta}$.

Let $\boldsymbol{w}_n = (w_{n1}, \ldots, w_{nV})$ where

$$w_{nj} = \begin{cases} 1 & \text{if the } n\text{th word in the document is the} \\ & j\text{th word in the vocabulary} \\ 0 & \text{otherwise} \end{cases}$$

As given in the generative model, $\boldsymbol{w}_n|\boldsymbol{Z}_n \sim mult(1, \phi_{1z_n}, \ldots, \phi_{Vz_n})$. Thus

$$f(\boldsymbol{w}_n|\boldsymbol{Z}_n) = \prod_{i=1}^{K} \prod_{j=1}^{V} \left(\phi_{ji}^{Z_{ni}}\right)^{w_{nj}}$$

and the likelihood of $\boldsymbol{w}|\boldsymbol{Z}$ is found as

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

where $\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N)$. Here ϕ_{ji} may be interpreted as the probability that word j is in the document given the topic allocation is i, $j = 1, \ldots, V$ and $i = 1, \ldots, K$.

The prior distribution of $\boldsymbol{\theta}$ is $Dir(\boldsymbol{\alpha})$ and for $\boldsymbol{\phi}$ is $Dir(\boldsymbol{\beta})$ thus these priors are proportional to

$$\begin{array}{rcl} \pi(\boldsymbol{\theta}|\boldsymbol{\alpha}) & \propto & \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \\ \pi(\boldsymbol{\phi}|\boldsymbol{\beta}) & \propto & \prod_{i=1}^{K} \prod_{j=1}^{V} \phi_{ji}^{\beta_{ji} - 1} \end{array}$$

The joint posterior distribution of the latent variable $\boldsymbol{Z}$, and parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is thus

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

2.1 Estimation Using Gibb's Sampling

We may develop fully conditional posteriors for Gibbs sampling. It is easily seen that

$$\begin{array}{rcl} \pi(\boldsymbol{\theta}|\boldsymbol{Z}, \boldsymbol{\phi}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) & \sim & Dir\left(\sum_{n=1}^{N} Z_{ni} + \alpha_i\right)_{i=1,\ldots,K} \\ \pi(\boldsymbol{\phi}|\boldsymbol{\theta}, \boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) & \sim & Dir(\left(\sum_{n=1}^{N} Z_{ni} w_{nj} + \beta_{ji}\right)_{j=1,\ldots,V,i=1,\ldots,K} \end{array}$$

The fully conditional posterior distribution for $\boldsymbol{Z}_n$ is found as

$$\begin{array}{rcl} \pi(\boldsymbol{Z}_n|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) & \propto & \prod_{i=1}^{k} \left(\theta_i^{Z_{ni}} \prod_{j=1}^{V} \left(\phi_{ji}^{w_{nj}}\right)^{Z_{ni}}\right) \\ & = & \prod_{i=1}^{K} \left(\theta_i \prod_{j=1}^{V} \phi_{ji}^{w_{nj}}\right)^{Z_{ni}} \end{array}$$

which has the form of a multinomial, $Mult\left(1, \theta_i \prod_{j=1}^{V} \phi_{ji}^{w_{nj}}\right)$.

To compute the marginal posterior distributions using Gibbs sampling, start with initial values of θ_i, $i = 1, \ldots, K$ and initial values of ϕ_{ji} for $i = 1, \ldots, K$ and $j = 1, \ldots, V$. Draw $\boldsymbol{Z}_n$ for $n = 1, \ldots, N$ then draw $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Continue this until convergence.

3 METHODS

Once the LDA model has been fit to a corpus it is important to assess how well the model fits. Traditionally, model fit can be assessed by randomly dividing the data set into a training set and a testing set. Blei et al. (2003) use 90% of the data for training purposes and the remaining 10% are used to assess model fit. Other means of cross validation such as leave-one-out cross-validation and m-fold cross validation can be used. Measures of goodness of fit can be computed on the trained and the tested data set for model comparison.

3.1 Perplexity

Traditionally, the perplexity of a held out data set (testing set or left out data set) has been used to evaluate the model. The perplexity is used by convention in language modeling and is given by

$$perplexity(D_{test}) = \exp\left(- \frac{\sum_{d=1}^{D} \log p(\boldsymbol{w}_d|\mathbf{Z}, \phi)}{\sum_{d=1}^{D} N_d} \right). \qquad (1)$$

The lower the perplexity score the better the general performance of the model and it is typically used to compare models. If perplexity is considered as a function of the number of topics, the model with minimum perplexity is considered the optimal model. Finding the minimum perplexity requires sensitivity analysis in order to compare the perplexitiy values with varying number of topics. Figure 2 shows the graph of a typical perplexity distribution as a function of k the number of topics. The graph in Figure 2 is separated into three areas where the perplexity graph has very different

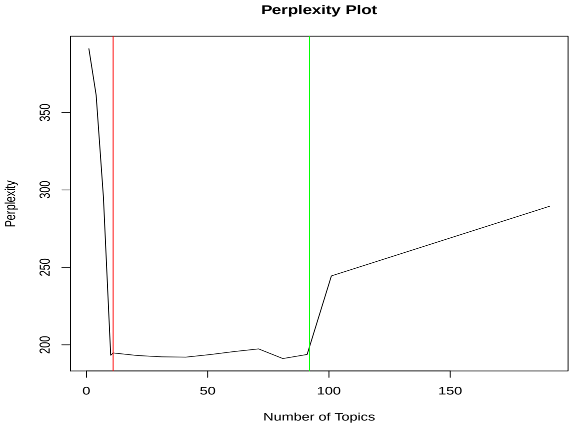

Fig. 2. Perplexity plot by Number of Topics

behavior. For low values of k, the number of topics, to the left of the first vertical line the perplexity drops very sharply. The high perplexity at lower k values is associated with underfitting and at higher k's is associated with overfitting.

The perplexity does not fully use the posterior distributions of the parameters that are output from the analysis and does not provide a means of testing for goodness of fit. Here we introduce frequentist methods that can be used to test the fitness of the model and Bayesian methods that use information from the posterior distributions of the parameters.

In assessing the fit of the model there are two measures of interest, each of which is useful in specific situations. Interest may be primarily in how well the derived model predicts words in a document, particularly for a new document in a corpus. Alternately, interest may be in how well the model predicts which topics are in a document, or how well the LDA model clusters the data. Frequentist and Bayesian measures of goodness of fit for both situations are derived.

3.2 Goodness of Fit of Words

The goodness of the model fit to words per document measures how well the model predicts the number of words from each possible word in the vocabulary that appear in the document. The actual numbers are observed in the document as $\sum_{n=1}^{N} w_{nj}$ for $j = 1, \ldots, V$.

To test the fit of the trained word distribution, we test the null hypothesis that words have been randomly assigned to topics. In the notation previously developed this is equivalent to testing $H_0 : \phi_{ji} = \phi_{\ell i} = 1/V$ for $i = 1, \ldots, K$, where ϕ_{ji} is interpreted as the probability that word j is in a document given it comes from topic i. Let M_i be the number of words assigned to topic i by the currently trained model for $i = 1, \ldots, K$ and W_{ji} be the number of times vocabulary word j comes from topic i. Assuming conditional independence of words given topic, $W_{1i}, \ldots, W_{Vi}$ follow a multinomial distribution, $(M_i, \phi_{ji}; j = 1, \ldots, K)$ for each topic i. Under the null hypothesis, H_0 the likelihood is

$$L(\boldsymbol{\phi}) = \prod_{i=1}^{K} \prod_{j=1}^{V} \left(\frac{1}{V}\right)^{W_{ji}} . \tag{2}$$

The maximum likelihood estimates of ϕ_{ji} are $\hat{\phi}_{ji} = W_{ji}/M_i$. The likelihood ratio function is then

$$\lambda_w = \prod_{i=1}^{K} \prod_{j=1}^{V} \left(\frac{M_i}{W_{ji}V}\right)^{W_{ji}} . \tag{3}$$

From this we get the log likelihood ratio test statistic as

$$T = -2\log\lambda_w = 2\sum_{i=1}^{K}\sum_{j=1}^{V} W_{ji}\log\left(\frac{W_{ji}}{e_i}\right) \tag{4}$$

where $e_i = M_i/V$ is the expected cell count under the null hypothesis. The test statistic T is referred to a χ^2_p for $p = K(V-1)$.

A Bayesian measure of goodness of fit of the word distribution can be derived using the log pointwise predictive density which is derived as follows. Recall that $\boldsymbol{w}$ is the vector of zeros and ones indicating which words are in a single document and we will use $\boldsymbol{w}_d$ to represent this vector for the dth document. The likelihood of $\boldsymbol{w}_d$ given the parameters was found to be

$$L(\boldsymbol{w}_d|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{n=1}^{N}\prod_{i=1}^{K}\prod_{j=1}^{V} (\phi_{ji})^{z_{ni}w_{nj}} . \tag{5}$$

The computed log pointwise predictive density for a corpus of documents of size D is found using draws from the posterior distributions derived from the LDA model. For $s = 1, \ldots, S$, draw values from the posteriors for $\boldsymbol{z}, \boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Label the sth draws as $\boldsymbol{z}^s, \boldsymbol{\theta}^s$ and $\boldsymbol{\phi}^s$. S is chosen to be large enough so that the posterior distributions have been fully captured. Use the parameters from the sth draw to evaluate 5, $L(\boldsymbol{w}|\boldsymbol{z}^s, \boldsymbol{\theta}^s, \boldsymbol{\phi}^s)$. The computed log pointwise predictive density, *lppd*, for the corpus is then found as

$$lppd = \sum_{d=1}^{D} \log\left(\frac{1}{S}\sum_{s=1}^{S} L(\boldsymbol{w}_d|\boldsymbol{z}^s, \boldsymbol{\theta}^s, \boldsymbol{\phi}^s)\right) \tag{6}$$

The $lppd$ as a measure of predictive accuracy of observed data overestimates the expected $lppd$ when it is evaluated on the data from which the model is fit. Thus if interest is in measuring the information criteria on the observed data set, an adjustment can be made to account for this bias. If a training and testing set is used to measure out-of sample prediction accuracy, this adjustment is not needed for the training set. An estimate of the effective number of parameters can be used to adjust the lppd in this case.

The large number of terms in the product in 5 used to estimate the log posterior predictive mean makes such a measure computationally infeasible for a large vocabulary and corpus. Instead the posterior probabilities from the trained topic model can be used to simulate observations from the generative process in order to assess how well the trained model can predict documents in the corpus. Output from topic modeling from varying software packages such as mallet and packages in R is in the form of posterior estimates of $\boldsymbol{\theta}|words$ and $\boldsymbol{\phi}|words, topics$. We can make use of these posterior estimates by using them to generate sets of *expected* documents which we can compare to existing documents as a measure of goodness of fit. For each word within each document choose a topic Z from a multinomial(1, $\boldsymbol{\theta}^d$). Using this topic, choose a word W from a multinomial (1, $\boldsymbol{\phi}_Z$). Compute a test statistic comparing the number of words per topic for each document generated from the posterior distributions to the observed word counts using a chi-square goodness of fit statistic. By generating a large number of samples, S, this way and averaging over the number of samples we get a monte carlo estimate of the mean goodness of fit of the generative posterior model. Large values of the monte carlo estimate compared with degrees of freedom of $V \times D$ indicate a poor fit of the trained generative model to the corpus.

3.3 Topic Goodness of Fit

The latent Dirichlet allocation model can be used to cluster observations (Zhao et al. 2014) within a corpus. For this type of application interest is in how well a model identifies topics. A frequentist test for topic goodness of fit tests the null hypothesis that topics are randomly assigned to documents. Specifically, the null hypothesis is $H_0 : \theta_i^{(d)} = \theta_\ell^{(d)} = 1/K$ for $i, \ell = 1, \ldots, K$ $i \neq \ell$ and for $d = 1, \ldots, D$. Here, as before, $\theta_i^{(d)}$ is the probability that topic i is included in document d. Let $Z_i^{(d)}$ be the number of words in document d that are from topic i. Then $(Z_1^{(d)}, \ldots, Z_K^{(d)})$ follow a multinomial distribution with $(N_d, \boldsymbol{\theta}^{(d)}$ for each document. Under H_0, the likelihood is maximized when $\hat{\theta}_i^{(d)} = 1/K$. The likelihood ratio function is then given by

$$\lambda_z = \prod_{d=1}^{D} \prod_{i=1}^{K} \left(\frac{N_d}{K Z_I^{(d)}} \right)^{Z_i^{(d)}} . \tag{7}$$

The log likelihood ratio test statistic is then

$$T = -2 \log(\lambda_z) = 2 \sum_{d=1}^{D} \sum_{i=1}^{K} Z_i^{(d)} \log \left(\frac{Z_i^{(d)}}{e_d} \right) \tag{8}$$

where $e_d = N_d/K$ is the expected topic counts under H_0. The test statistic, T has an approximate χ_p^2 under H_0 with $p = (K-1)D$.

A Bayesian measure of goodness of topic clustering can be derived using the equation

$$p(\boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\phi}) \propto p(\boldsymbol{w}|\boldsymbol{z}_i, \boldsymbol{\phi}) P(\boldsymbol{z}_i|\boldsymbol{\theta}) = \prod_{n=1}^{N} \theta_i^{z_{ni}} \prod_{j=1}^{V} \phi_{ji}^{z_{ni} w_{nj}} \tag{9}$$

for a single document. The computed log pointwise predictive density is then proportional to

$$lppd = \sum_{d=1}^{D} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(\boldsymbol{z}_i|\boldsymbol{w}_d, \boldsymbol{\theta}^s, \boldsymbol{\phi}^s) \right) \tag{10}$$

where $p(\boldsymbol{z}_i|\boldsymbol{w}_d, \boldsymbol{\theta}^s, \boldsymbol{\phi}^s)$ is 9 evaluated at the sth draw from the posterior distributions of the parameters for $s = 1, \ldots, S$ and the constant of proportionality is not a function of $\boldsymbol{Z}$.

As for the words per distribution goodness of fit, if the *lppd* is computed on the observed data set an adjustment should be made for overfitting by subtracting the effective number of parameters estimated.

As in the case of goodness of fit for words, the *lppd* for the topics also suffers computationally due to the large number of terms in the product. A goodness of fit based on the generative process cannot be used for the topic fit since the topics are latent and not observed.

4 RESULTS

The developed goodness of fit measures are illustrated on two different datasets. The first data set is retrieved from the publicly available SIDER2 database (http://sideeffects.embl.de) discussed in Kuhn *et al.*, 2010. The dataset includes 996 drugs (documents, D = 996) with 3,034 side effects (words, V=3034) for the drugs. The dataset can be envisioned as a 996 × 4500 matrix of 1's and 0's with one indicating the side effect is included in the drug profile. Topic models with different number of topics were fit to the data. For this dataset Zhao *et al.*, 2015, using a heuristic approach based on perplexity scores found 50 to be the optimal number of topics for this data. The goodness of fit test of H_0 : *Topics are randomly assigned within documents* and H_0: *Words are randomly assigned to topics* for various values of K are shown in Table 1 below.

Table 1 shows 1/2 the log likelihood ratio values for measuring both topic and word goodness of fit. The tests all are highly significant in their rejection of the null hypotheses. The degrees of freedom when testing for topic GOF are $(K-1)D$ ranging from $8,962$ to $98,604$ as K goes from 10 to 100. For the word GOF the degrees of freedom are $K(V-1)$ with values ranging from 30330 to 303300. For all values of K the trained topic models are clearly not assigning either words or topics randomly using the goodness of fit tests proposed. In Table 1 the proportions of individual tests that exceeded 0.05 is included for both topic GOF and word GOF. For topics, the proportion of individual tests that exceed 0.05 is the number of documents with p-values from a log likelihood ratio test statistic with $K-1$ degrees of freedom that exceeded 0.05. This proportion is seen to spike at $K = 20$ and fall to 0.0060 at $K = 40$, increasing with K afterwards. This demonstrates that for the SIDER2 data, topic goodness of fit proportions per document can be useful in choosing an optimum value of K. For this data set $K = 40$ is preferred according to this criteria. This is not the same optimum value as found in Zhao *et al.*, 2015 but is in the middle range of their perplexity graph. For the word goodness of fit, individual log-likelihood ratio tests are computed within each topic (degrees of freedom $V-1$) and the proportion of individual tests whose p-values exceed 0.05 are shown in Table 1. The proportion of individual word tests with large p-values is 0 until a large number of topics is proposed. When $K = 100$ in the current example, the proportion of individual tests with large p-values becomes 0.1100. This is evidence of overfitting by the inclusion of too many topics.

Table 1. Goodness of fit measures for SIDER data

	Topic		Word	
K	GOF	Prop > 0.05	GOF	Prop > 0.05
10	61,483	0.0291	223,806	0.0000
20	83,902	0.1757	268,853	0.0000
30	87,235	0.0833	301,735	0.0000
40	88,369	0.0060	329,196	0.0000
50	93,692	0.1546	343,419	0.0000
100	104,864	0.3042	400,701	0.1100

For the SIDER2 data the proposed goodness of fit of the Bayesian generative model resulted in values of the test statistic

$$\sum_{d=1}^{D} \frac{(GD^{(d)} - OB^{(d)})^2}{GD^{(d)}} \tag{11}$$

near $573,000$ for all values of K, demonstrating that it is not dependent on the number of topics chosen. Here $GD^{(d)}$ is the document generated by the trained LDA model. It is a $V \times 1$ vector with jth element consisting of the number of times the jth word in the vocabulary occurs in generated document d. The $K \times 1$ vector $OB^{(d)}$ is the vector of observed word counts of the dth document. The notation above refers to summing the square of the differences between observed and generated divided by the generated over all words in the vocabulary and summing these values over all documents. The degrees of freedom of $V * D$ was $3,021,864$ indicating no lack of fit of the generative model to this dataset.

The second dataset was created by Zhao *et al.*, 2015 by retrieving the abstracts of papers published in the IEEE Transactions on Computational Biology and Bioinformatics (TCBB) from the PubMed database. The dataset was comprised of all the abstracts of (D =) 885 papers published in TCBB from 2004 to 2013. The vocabulary consisted of $V = 5004$ words once stopwords consisting of common English adverbs, conjunctions, pronouns and prepositions were removed. Table 2 gives the results of the topic and word GOF tests for the TCBB data.

For this data, all of the tests for random assignment of topic or words resulted in rejection of the null hypothesis except for the case where $K = 100$ when testing for random word per topic

Table 2. Goodness of fit measures for TCBB data

	Topic		Word	
K	GOF	Prop > 0.05	GOF	Prop > 0.05
10	45,114	0.0000	103,077	0.0000
20	60,800	0.0023	122,648	0.0000
30	65,298	0.0056	136,956	0.3670
40	66,168	0.0113	148,966	0.4750
50	71,645	0.0294	155,191	0.5600
100	78,474	0.1571	184,943	0.8300

assignment. The degrees of freedom for this test, $K(V-1)$ was equal to $500,300$ and the observed value of the log likelihood ratio test statistic was $2 * 184,943$ resulting in a p-value of 1. This indicates that 100 topics is overfitting.

Zhao *et al.*, 2015 found the optimum number of topics for this data set to be K = 40. In our analysis at $K = 40$ the proportion of individual tests of random assignment by document that had p-values over 0.05 in the topic goodness of fit was 0.0113. The proportion of individual tests of random assignment of words to topics that had p-values over 0.05 in the word goodness of fit test was 0.4750. So in nearly half of the topics at $K = 40$ the distribution of words was no better than randomly assigned. This proportion was zero for $K = 10$ and $K = 20$ and jumped sharply to 0.3670 at $K = 30$. The goodness of fit proportions of individual tests provides evidence that a smaller number of topics is preferred for this data. Zhao *et al.*, 2015 discuss the preferred choice of K using a heuristic approach and judge the resulting clustering of words as the degree to which topic themes were distinct and cohesive. Their model with 20 topics had some clusters with distinct and salient and some without these properties. Specifically the 20 topic model tended to combine themes that would be better separated and to produce some topics that are less specific than the forty topic model. Parsimony and meaningful assignment of topics should be balanced to select the optimum number of topics. Word clouds for the TBCC data with differing number of topics are included in Zhao *et al.*, 2015.

The value of the goodness of fit statistic of the Bayesian generative model for the TCBB data was $156,787$ for $K = 10$ and decreased steadily to $113,772$ when $K = 100$. The degrees of freedom for the generative goodness of fit statistic $V \times D$ was $4,428,540$. The observed values indicate no lack of fit of the generative model.

5 CONCLUSION

The goodness of fit measures developed here are a useful tool in evaluating trained LDA topic models. A test of whether topics are randomly assigned within document was derived. In addition, the proportion of documents which would fail to reject the random assignment hypothesis is a good indication of whether the model derives meaningful topics. A test for whether words are randomly assigned to topics was also derived. The proportion of topics with random word assignment provides information on whether the trained model has meaningful word assignments. Bayesian methods based on log predictive posterior distributions were derived but found to be impractical for comparing models in this context. Instead an estimate of how well the model simulates the assumed generative process is developed. The trained model is used to simulate a generated document and the degree that the generated model fits the observed documents is a measure of how well the assumption of a generative model is supported by the corpus.

The methods developed were applied to two datasets. Results were compared with the heuristic approach of Zhao *et al.*, 2015. The proposed tests of random assignment and of the generative process provided a means of determining an acceptable value for the unknown number of topics. The results using the methods proposed did not agree with the results of Zhao *et al.*, 2015, but supported a more parsimonious approach to selected the best number of topics.

Funding: This research 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 EM, Blei DM, Feingerg SE, Xing EP (2008) bf Mixed Membership Stochastic Blockmodels,*Journal of Machine Learning Research*, **9**, 1981-2014.

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

Blei, DM and Lafferty, JD (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, *Journal of Machine Learning Research*, **3**, 993-1022.

Coelho LP, Peng T, Murphy RF (2010) Quantifying the Distribution of Probes between Subcellular Locations Using Unsupervised Pattern Unmixing,*Bioinformatics*, **26**(12): 7–12.

Griffiths TL, Steyvers M (2004) Finding Scientific Topics, *Proceedings of the National Academy of Sciences of the United States of America*, **101**(suppl.): 5228–5235.

Hofmann T (2001) Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, **42**(1-2): 177–196.

Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P (2010) A Side Effect Resource to Caputure Phenotypic Effects of Drugs,*Molecular systems biology*, **6**: 343.

Rogers S, Girolami M, Campbell C, Breitling R (2005) The Latent Process Decomposition of cDNA Microarray Data Sets,*IEEE?ACM transactions on computational biology and bioinformatics/IEEE,ACM* , **2**(2)L 143–156.

Shivashankar S, Srivathsan S, Ravindran B, Tendulkar AV (2011) Multi-View Methods for Protein Structure Comparison Using Latent Dirichlet Allocation,*Bioinformatics*, **27**(13): 161–168.

Zhao W, Chen JJ, Liu Z, Ge W, Ding Y, Zou W (2015) A Heuristic Approach to Determine an Optimal Number of Topics in Topic Modeling, *submitted.*

Zhao W, Zou W, Chen JJ (2014) Topic modeling for cluster analysis of large biological and medical datasets, *BMC Bioinformatics*, **15 Suppl 11**:S11.