A Coefficient of Determination for Probabilistic Topic Models

Tommy Jones
Department of Computational and Data Sciences
George Mason University
Fairfax, VA
jones.thos.w@gmail.com

Abstract—This document proposes a new (old) metric for evaluating goodness of fit in topic models, the coefficient of determination, or R^2. Within the context of topic modeling, R^2 has the same interpretation that it does when used in a broader class of statistical models. Reporting R^2 with topic models addresses two current problems in topic modeling: a lack of standard cross-contextual evaluation metrics for topic modeling and ease of communication with lay audiences. This paper proposes that R^2 should be reported as a standard metric when constructing topic models.

Introduction

According to an often-quoted but never cited definition, "the goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question."[1] Goodness of fit measures vary with the goals of those constructing the statistical model. Inferential goals may emphasize in-sample fit while predictive goals may emphasize out-of-sample fit. Prior information may be included in the goodness of fit measure for Bayesian models, or it may not. Goodness of fit measures may include methods to correct for model overfitting. In short, goodness of fit measures the performance of a statistical model against the ground truth of observed data. Fitting the data well is generally a necessary—though not sufficient—condition for trust in a statistical model, whatever its goals.

Of course, goodness of fit is only one concern in statistical modeling. Researchers may trade some goodness of fit for ease of interpretation. For example, many accurate and robust predictive models are non-parametric "black boxes", making inference difficult. Researchers may trade off some goodness of fit for interpretability by selecting a more restrictive parametric model. If the emphasis is on prediction, researchers may trade some in-sample fit for predictive robustness. This is cited as one motivation for using the Bayesian latent Dirichlet allocation (LDA) topic model over the frequentist probabilistic latent semantic analysis (pLSA). It is alleged that pLSA tends to over fit its training sample, making its estimates fragile to the introduction of new data. (Blei, Ng, and Jordan 2003) Under certain conditions, pLSA and LDA are equivalent models, however. Girolami and Kabán (2003)

Goodness of fit manifests itself in topic modeling through word frequencies. It is a common misconception that topic models are fully-unsupervised methods. If true, this would mean that no outcomes exist upon which to compare a model's fitted values. However, topic models are ultimately generative models of word frequencies. The expected value of a document under a topic model are given by the expected value of a multinomial random variable. The outcomes, then, are the word frequencies themselves. Most goodness of fit measures in topic modeling are restricted to in-sample fit. However, some out-of-sample measures have been developed. Buntine (2009)

Probabilistic Topic Models

Probabilistic topic models are a family of stochastic models for estimating abstract "topics" in a set of documents. Many methods have been developed to provide a flexible family of topic models. Some include frequently available metadata about documents, such as the time of the publication [CITE DYNAMIC TOPIC MODELS], or the author and location of the publication [CITE]. Most probabilistic topic models are Bayesian, though probabilistic latent semantic analysis (pLSA) is frequentist. [CITE] Without loss of generality, all probabilistic topic models model the document-generating process as a mixture of multinomial distributions.[2] Probabilistic topic models estimate parameters of an idealized stochastic process for how words get on the page. Instead of writing full, syntatictically-coherent, sentences, the author samples a topic from a multinomial distribution and then given the topic samples a word. The process for a single draw of the n-th word for the d-th document, $w_{d,n}$, is

1. Sample $z_{d,n} \sim \text{Multinomial}(1, \boldsymbol{\theta}_d)$
2. Sample $w_{d,n} \sim \text{Multinomial}(1, \boldsymbol{\phi}_{z_{d,n}})$

The variable $z_{d,n}$ is latent. The author repeats this process Nd times until the document is "complete". For

[1] This quote appears verbatim on Wikipedia and countless books, papers, and websites.

[2] The terms "multinomial" and "categorical" are often used interchangeably in the topic modeling literature.

a corpus of D documents, V unique tokens, and K latent topics, the goal is to estimate two matrices: $\mathbf{\Theta}$ and $\mathbf{\Phi}$. The d-th row of $\mathbf{\Theta}$ comprises $\boldsymbol{\theta}_d$, above. And the k-th row of $\mathbf{\Phi}$ comprises $\boldsymbol{\phi}_k$.[3] The document term matrix (DTM)—$\mathbf{Y}$—can be thought of as the result of repeated sampling from $\mathbf{\Theta}$ and $\mathbf{\Phi}$. In expectation we have the following relationship:

$$E(\mathbf{Y}) = \mathbf{n} \odot \Theta \cdot \Phi \tag{1}$$

Above, $\mathbf{n}$ is a D-length vector whose d-th entry is the number of terms in the d-th document and $\odot$ denotes elementwise multiplication.

Evaluation Metrics for Topic Models

The primary goodness of fit measures in topic modeling are likelihood methods. Likelihoods, generally the log likelihood, are naturally obtained from probabilistic topic models. Likelihoods may contain prior information, as is often the case with Bayesian models. If prior information is unknown or undesired, researchers may calculate the likelihood using only estimated parameters. Researchers have used likelihoods to select the number of topics[CITE], compare priors[CITE], or otherwise evaluate the efficacy of different modeling procedures. [CITE] [CITE] A popular likelihood method for evaluating out-of-sample fit is called perplexity. Perplexity measures a transformation of the likelihood of the held-out words conditional on the trained model.

Non-likelihood measures are less-commonly used. Some of these methods bridge the gap between interpretation and goodness of fit. One such method is the intruder test.[CITE] The intruder test uses human judgment. Judges are shown a few high-probability words in a topic, with one low-probability word mixed in. Judges must find the low-probability word, the intruder. They then repeat the procedure with documents instead of words. A good topic model should allow judges to easily detect the intruders. Coherence is a measure attempting to approximate the results of intruder tests in an automated fashion. [CITE] The coherence of a topic is a sum of log ratios of word co-occurance across documents. Researchers have used precision and recall methods such as the area under a ROC curve (AUC) on topically-tagged corpora.[CITE] The most prevalent topic in each document is taken as a document's topical classification.

Though useful, current evaluation metrics in topic modeling are difficult to interpret, are inappropriate for use in topic modeling, or are cannot be produced easily. Intruder tests are time consuming and costly, making intruder tests infeasible to conduct regularly. Coherence is not primarily a goodness of fit measure. AUC is nicely bound between 0 and 1, with 0.5 representing "as good as random guessing". Yet precision and recall measures mis-represent topic models as binary classifiers. This misrepresentation ignores one fundamental motivation for using topic models: allowing documents to contain multiple topics. Precision and recall measures also require substantial subjective judgement. Researchers must examine the high-probability words in a topic and decide that it does or does not correspond to the corpus tags.

Likelihoods have an intuitive definition: the probability of observing the training data if the model is true. Yet properties of the underlying corpus influence the scale of the likelihood function. Adding more documents, having a larger vocabulary, and even having longer documents all reduce the likelihood. Likelihoods of multiple models on the same corpus can be compared. (Researchers often do this to help select the number of topics for a final model.)[CITE] Topic models on different corpora cannot be compared, however. One corpus may have 1,000 documents and 5,000 tokens, while another may have 10,000 documents and 25,000 tokens. The likelihood of a model on the latter corpus will be much smaller than a model on the former. Yet this does not indicate the model on the latter corpus is a worse fit; the likelihood function is simply on a different scale. Perplexity is a transformation of the likelihood for out-of-sample documents. The transformation makes perplexity less-intuitively interpreted than a raw likelihood. Perplexity's scale is influenced by the same factors as the likelihood.

The Coefficient of Determination: R^2

The coefficient of determination is a popular, intuitive, and easily-interpretable goodness of fit measure. The coefficient of determination, denoted R^2, is most common in ordinary least squares (OLS) regression. However, researchers have developed R^2 and several pseudo R^2 measures for many classes of statistical models. The largest value of R^2 is 1, indicating a model fits the data perfectly. The formal definition of R^2 (below) is interpreted—without loss of generality—as the proportion of variability in the data that is explained by the model. For linear models with outcomes in $\mathbb{R}_1$, R^2 is bound between 0 and 1 and is the proportion of variance in the data explained by the model.[CITE] Even outside of the context of a linear model, R^2 retains its maximum of 1 and its interpretation as the proportion of explained variability. Negative values of R^2 are possible for non-linear models or models in $\mathbb{R}_m$ where $m > 1$. These negative values indicate that simply guessing the mean outcome is a better fit than the model.[4]

The Standard Definition of R^2

For a model, f, of outcome variable, y, where there are n observations, R^2 is derived from the following:

[3] LDA estimates these parameters by placing Dirichlet priors on θ_d and ϕ_k.

[4] In this author's experience, negative values indicate an error in one's code, rather than simply due to an exceptionally poor fitting model.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2}$$

$$SS_{tot.} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3}$$

$$SS_{resid.} = \sum_{i=1}^{n} (f_i - y_i)^2 \tag{4}$$

Thus, the standard definition of R^2 is a ratio of summed squared errors.

$$R^2 \equiv 1 - \frac{SS_{resid.}}{SS_{tot.}} \tag{5}$$

A Geometric Interpretation of R^2

R^2 has a geometric interpretation as well. $SS_{tot.}$ is the total squared-euclidean distance from each y_i to the mean outcome, $\bar{y}$. Then $SS_{resid.}$ is the total squared-euclidean distance from each y_i to its predicted value under the model, f_i. Recall that for any two points $\mathbf{p}, \mathbf{q} \in \mathbb{R}_m$

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{j=1}^{m} (p_j - q_j)^2} \tag{6}$$

where $d(\mathbf{p}, \mathbf{q})$ denotes the euclidean distance between $\mathbf{p}$ and $\mathbf{q}$. R^2 is often taught in the context of OLS where $y_i, f_i \in \mathbb{R}_1$. In that case, $d(y_i, f_i) = \sqrt{(y_i - f_i)^2}$; by extension $d(y_i, \bar{y}) = \sqrt{(y_i - \bar{y})^2}$. In the multidimensional case where $\mathbf{y}_i, \mathbf{f}_i \in \mathbb{R}_m; m > 1$, then $\bar{\mathbf{y}} \in \mathbb{R}_m$ represents the point at the center of $\mathbf{y}$ in $\mathbb{R}_m$.[5]

We can rewrite R^2 using the relationships above.

$$\bar{\mathbf{y}}_v = \frac{1}{n}\sum_{i=1}^{n} y_{i,v}, \text{ the } v\text{-th entry of the vector } \bar{\mathbf{y}} \tag{7}$$

$$SS_{tot.} = \sum_{i=1}^{n} d(\mathbf{y}_i, \bar{\mathbf{y}})^2 \tag{8}$$

$$SS_{resid.} = \sum_{i=1}^{n} d(\mathbf{y}_i, \mathbf{f}_i)^2 \tag{9}$$

$$\Rightarrow R^2 = 1 - \frac{\sum_{i=1}^{n} d(y_i, f_i)^2}{\sum_{i=1}^{n} d(y_i, \bar{y})^2} \tag{10}$$

Figure [REFERENCE FIGURE] visualizes the geometric interpretation of R^2 for outcomes in $\mathbb{R}_2$. The left image represents $SS_{tot.}$: the red dots are data points(y_i); the black dot is the mean ($\bar{y}$); the line segments represent the euclidean distance from each y_i to $\bar{y}$. $SS_{tot.}$ is obtained by squaring the length of each line segment and then adding them together. The right image represents $SS_{resid.}$: the blue dots are the fitted values under the model (f_i); the line segments represent the euclidean distance from each f_i to its corresponding y_i. $SS_{resid.}$ is obtained by squaring the length of each line segment and then adding them together.

[5] In the one-dimensional case, where $y_i, f_i \in \mathbb{R}_1$, $SS_{resid.}$ can be considered the squared-euclidean distance between the n-dimensional vectors y and f. However, this relationship does not hold when $\mathbf{y}_i, \mathbf{f}_i \in \mathbb{R}_m; m > 1$.

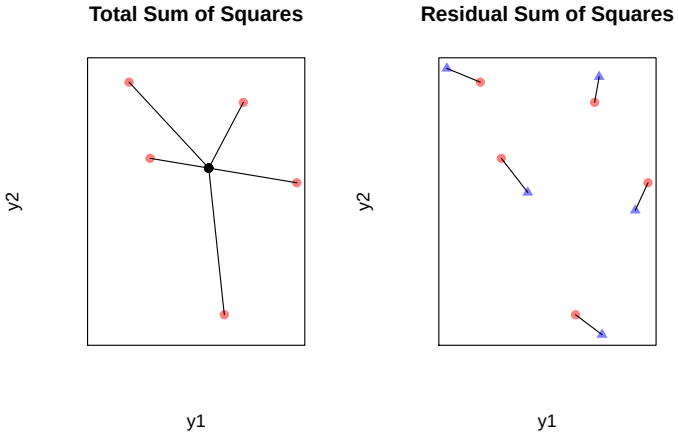

Fig. 1. Visualizing the geometric interpretation of R-squared: corresponds to an R-squared of 0.87

The geometric interpretation of R^2 is similar to the "explained-variance" interpretation. When $SS_{resid.} = 0$, then the model is a perfect fit for the data and $R^2 = 1$. If $SS_{resid.} = SS_{tot.}$, then $R^2 = 0$ and the model is no better than just guessing $\bar{y}$. When $0 < SS_{resid} < SS_{tot}$, then the model is a better fit for the data than a naive guess of $\bar{y}$. In a non-linear or multi-dimensional model, it is possible for $SS_{resid.} > SS_{tot.}$. In this case, R^2 is negative, and guessing $\bar{y}$ is better than using the model.

Extending R^2 to Topic Models

An R^2 for topic models follows from the geometric interpretation of R^2. For a document, d, the observed value, y_d, is a vector of integers counting the number of times each token appears in N_d draws. The document's fitted value under the model follows that y_d represents the outcome of a multinomial random variable. The fitted value is $f_d = N_d \cdot \theta_d \cdot \Phi^T$. The center of the documents in the corpus, $\bar{y}$, is obtained by averaging the occurrence of each token across all documents. From this we obtain R^2.

$$\bar{y}_v = \frac{1}{D}\sum_{d=1}^{D} y_{d,v} \tag{11}$$

$$SS_{tot.} = \sum_{d=1}^{D} d(y_d, \bar{y})^2 \tag{12}$$

$$SS_{resid.} = \sum_{d=1}^{D} d(y_d, f_d)^2 \tag{13}$$

$$R^2 \equiv 1 - \frac{SS_{resid.}}{SS_{tot.}} \tag{14}$$

Pseudo Coefficients of Variation

Several pseudo coefficients of variation have been developed for models where the traditional R^2 is inappropriate. Some of these, such as the Cox and Snell's R^2 [?] or McFadden's R^2 [?] may apply to topic models. As pointed out by UCLA's Institute for Digital Research and Education,[?]

> These are 'pseudo' R-squareds because they look like R-squared in the sense that they are on a similar scale, ranging from 0 to 1 (though some pseudo R-squareds never achieve 0 or 1) with higher values indicating better model fit, but they cannot be interpreted as one would interpret an OLS R-squared and different pseudo R-squareds can arrive at very different values.

The empirical section of this paper calculates an uncorrected McFadden's R^2 for topic models to compare to the standard (non-pseudo) R^2. McFadden's R^2 is defined as

$$R^2_{Mc} \equiv 1 - \frac{ln(L_{full})}{ln(L_{restricted})} \tag{15}$$

$$\tag{16}$$

where $ L_{full} $ is the estimated likelihood of the data under the model and $ L_{restricted} $ is the estimated likelihood of the data free of the model. In the context of OLS, the restricted model is a regression with only an intercept term. For other types of model (such as topic models), care should be taken in selecting what "free of the model" means.

Empirical Evaluation of Topic Model R^2

This paper performs three analyses to empirically evaluate R^2 for topic models. Two analyses use Monte Carlo-simulated corpora and one uses a corpus of grants awarded through the Department of Health and Human Services. This latter corpus was obtained from the National Institutes of Health NIH ExPORTER database. [?] The first analysis uses simulated corpora to observe how the properties of training corpora influence R^2. The second analysis uses these same simulated corpora to compare R^2 as commonly-defined to McFadden's R^2 in the context of topic modeling. The final analysis compares the R^2 values of various models constructed on the NIH corpus.

Monte Carlo-Simulated Corpora

It is possible to simulate corpora that are statistically-consistent to human-generated language. St. Thomas and Jones derive this method in a working paper recently submitted for peer review. [?] This method generates a corpus of D documents, V tokens, and K topics through the following stochastic process.

$$\beta_{k,v} \sim Pareto(1, \gamma) \qquad k \in \{1, \ldots, K\} \tag{17}$$

$$\phi_k \sim Dirichlet_V(\beta_k) \qquad k \in \{1, \ldots, K\} \tag{18}$$

$$\theta_d \sim Dirichlet_K(\alpha) \qquad d \in \{1, \ldots, D\} \tag{19}$$

$$N_d \sim Poisson(\lambda) \qquad d \in \{1, \ldots, D\} \tag{20}$$

$$z_{d,n} \sim Categorical_K(\theta_d) \qquad d \in \{1, \ldots, D\}, n \in \{1, \ldots, N_d\} \tag{21}$$

$$v_{k,n} \sim Categorical_V(\phi_k) \qquad k \in \{1, \ldots, K\}, n \in \{1, \ldots, N_d\} \tag{22}$$

$$\tag{23}$$

The words for document d are populated by sampling with replacement from $z_{d,n}$ and $v_{k,n}$ for N_d iterations. The parameters V, K, and λ may be varied to adjust the corpus properties for the number of tokens, topics, and average document length respectively. β_k is drawn from a Pareto distribution to generate corpora consistent with Zipf's law. The parameter of the Pareto distribution, γ, may be estimated from an existing corpus; St. Thomas and Jones set $\gamma = 1.1$. The index, v, of each token is permuted for each β_k to ensure heterogeneity of the simulated topics.

Latent Dirichlet Allocation

This paper uses Latent Dirichlet Allocation (LDA) and collapsed Gibbs sampling to estimate topic models from simulated or actual data. LDA, possibly the most popular topic model, is a Bayesian hierarchical model that places Dirichlet priors on $\theta_d, \forall d$ and $\phi_k, \forall k$:

$$\begin{aligned} \phi_k &\sim Dirichlet_V(\beta) & k &\in \{1, \ldots, K\} \\ \theta_d &\sim Dirichlet_K(\alpha) & d &\in \{1, \ldots, D\} \end{aligned}$$

A limitation with LDA is that the number of topics in the corpus must be specified a priori where no prior knowledge often exists. Compounding the problem, LDA's posteriors are inconsistent with respect to the number of topics. [?] Nevertheless, LDA's empirical utility has been overwhelmingly demonstrated, having been used since 2002. For LDA models estimated in this paper, symmetric parameters are used for both prior distributions. Each entry of α is 0.1; each entry of β is 0.05.

Comparing R^2 to McFadden's R^2

McFadden's pseudo R^2 is calculated for all simulated corpora for comparison to the standard R^2. McFadden's R^2 is a ratio of likelihoods. Various methods exist for calculating likelihoods of topic models. [?] Most of these methods have a Bayesian perspective and incorporate

prior information. Not all topic models are Bayesian, however. And in the case of simulated corpora, the exact data-generating parameters are known a priori. As a result, likelihoods calculated for this paper follow the simplest definition: the probability of observing the generated data, given the (known or estimated) multinomial parameters of the model. The model-free" likelihood assumes that each document is generated by drawing from a single multinomial distribution. The parameters of thismodel-free" distribution are proportional to the frequency of each token in the data.

Empirical Properties of R^2 for Topic Models

The R^2 for topic models has the following empirical properties: It is bound between $-\infty$ and 1. R^2 is invariant to the true number of topics in the corpus. R^2 increases with the estimated number of topics using LDA. This indicates a risk of model over fit in selecting too many topics (discussed in more detail in the next section). R^2 decreases slightly as the vocabulary size of the corpus increases. There are large increases in R^2 as the average document length increases. Figures depicting these properties are in the Appendix.

Most of these empirical properties were obtained by using simulated corpora, as described earlier in the paper, with one exception. The default parameter settings for Monte Carlo simulation are $K = 50$ topics, $D = 2,000$ documents, $V = 5,000$ tokens, and $\lambda = 500$ for the average document length, which is distributed $Poisson(\lambda)$. Each parameter was varied, holding other parameters constant. In each case, Monte Carlo simulation generates a document term matrix for the corpus while the parameters for generating the documents are known. R^2 is calculated for each simulated corpus, using the population parameters. These R^2 metrics represent a best-case scenario, avoiding misspecification and other pathologies present with topic model estimation algorithms.

McFadden's pseudo R^2 is also calculated for these simulated data. McFadden's R^2 is uniformly lower than the standard R^2. McFadden's R^2 is subject to the common problem of many pseudo R^2 measures; its true upper bound is less than one. This makes sense. If McFadden's R^2 were to equal 1, then the likelihood of the data would be 1 which is impossible. It is never the case that a likelihood will equal 1. Given the scale of linguistic data, the likelihood will always be significantly less than 1. Therefore, a scale correction measure is needed, making McFadden's R^2 more complicated.

McFadden's R^2 increases with the number of true topics; this is problematic. When the scale of an R^2 varies with known properties of the data, such as the number of documents, vocabulary size, average document length, etc. scale correction measures are possible. However, when the metric varies with an unknown property, such as the number of latent topics, then a scale correction is not possible. For this reason alone, McFadden's R^2 is undesirable. Other properties of McFadden's R^2 are consistent with the properties of the standard R^2

These properties of R^2 compel a change in focus for the topic modeling research community. Document length is an important factor in model fit whereas the number of documents is not. When choosing between fitting a topic model on short abstracts, full-page documents, or multi-page papers, it appears that more text is better. Second, since model fit is invariant for corpora over 1,000 documents (our lower bound for simulation), sampling from large corpora should yield reasonable estimates of population parameters. The topic modeling community has heretofore focused on scalability on large corpora of hundreds-of-thousands to millions of documents. Little attention has been paid to document length, however. These results, if robust, indicate that focus should move away from larger corpora and towards lengthier documents.

Comparison of Simulated Data with the NIH Corpus

R^2 increases with the estimated number of topics using LDA. This indicates a risk of model over fit with respect to the number of estimated topics. To evaluate the effect of R^2 on the number of estimated topics, LDA models were fit to two corpora. The first corpus is simulated, using the parameter defaults: $K = 50$, $D = 2,000$, $V = 5,000$, and $\lambda = 500$. The second corpus is on the abstracts of 10,000 randomly-sampled research grants for fiscal year 2014 in the National Institutes of Health's ExPORTER database. In both cases, LDA models are fit to the data estimating a range of K. For each model, R^2 and the log likelihood are calculated. The known parameters for this NIH corpus are $D = 10,000$, $V = 21,064$, and the median document length is 368 words. This vocabulary has standard stop words removed, includes bigrams as tokens, and excludes tokens that appear in fewer than 5 documents or more than half of the corpus.

The same likelihood calculation is used here, as described above. This likelihood calculation excludes prior information typically used when calculating the likelihood of an LDA model. This is perhaps not the optimal method for calculating likelihoods when using a Bayesian model. However, this prior information is excluded here for two reasons. First, it is consistent with the method used for calculating McFadden's R^2 earlier in the paper. Second, this log likelihood can be calculated exactly. LDA's likelihood is contained within an intractable integral. Various approximations have been developed. However, there is some risk that a comparison of an approximated likelihood to R^2 may be biased by the approximation method.

[FIGURE HERE]

Figure ˜?? depicts R^2 and the log likelihood over a range of estimated K for both corpora. The left image corresponds to the simulated corpus. The right image corresponds to the NIH corpus. From both images, we see that R^2 and the log likelihood both increase with the number of estimated topics at approximately the same

rates, proportional to their scales. The scale of R^2 is comparable between the simulated corpus and the NIH corpus, in spite of their differences in size. The scale of log likelihoods between the corpora differ by orders of magnitude. In the case of the simulated corpus, we know that the true number of topics is 50. However, this is not clear from observing R^2 or the log likelihood. R^2 does appear to flatten between 55 and 60 topics for the simulated corpus while it continues to increase for the NIH corpus. Nevertheless, it does not appear that R^2 or this "raw" likelihood can help find the true number of topics.[6]

In [?], Chang et al. observe that there may be a trade off between model fit and human interpretation. Specifically, they find that humans can more-easily interpret models with fewer topics. This may be true. As discussed in an earlier section, non-R^2 goodness of fit measures are not readily comparable across corpora and models. Because R^2 is comparable, it is now possible to quantify how much goodness-of-fit is lost when K is lowered. This does not address any issues arising from a pathological misspecification of the model, an "incorrect" K, for example. In the case of the NIH corpus, R^2 goes from 0.5 to 0.4 when the number of estimated topics is lowered from 200 to 100.

Conclusion

R^2 has many advantages over standard goodness of fit measures commonly-used in topic modeling. Current goodness of fit measures are difficult to interpret, compare across corpora, and explain to lay audiences. R^2 does not have any of these issues. Its scale is effectively bound between 0 and 1, as negative values (though possible) are rare and indicate extreme model misspecification. R^2 may be used to compare models of different corpora, if necessary. Scientifically-literate lay audiences are almost uniformly familiar with R^2 in the context of linear regression; the topic model R^2 has a similar interpretation, making it an intuitive measure.

The standard (geometric) interpretation of R^2 is preferred to McFadden's pseudo R^2. The effective upper bound for McFadden's R^2 is considerably smaller than 1. A scale correction measure is needed. Also, it is debatable which likelihood calculation(s) are most appropriate. These issues make McFadden's R^2 complicated and subjective. However, a large motivation for deriving a topic model R^2 is to remove the complications that currently hinder evaluating and communicating about topic models. Most problematic, McFadden's R^2 varies with the number of true topics in the data. It is therefore unreliable in practice where the true number of topics in unknown.

A result of this paper indicates that further study is needed on the relationship between document length, the number of documents in a corpus, and vocabulary size. Results reported in this paper demonstrate that document length is a considerable factor in model fit, whereas the number of documents (above 1,000) is not. If robust, this result indicates that the topic modeling community may need to change focus away from scaling estimation algorithms for large corpora. Instead, more effort should be put towards obtaining high-quality data. Also, studying the relationship between these parameters, along with the number of topics, may facilitate the development of an adjusted R^2, guarding against model over fit.

Lack of consistent evaluation metrics has limited the use of topic models as a mature statistical method. The development of an R^2 for topic modeling is no silver bullet. However it represents a step towards establishing consistency and rigour in topic modeling. This paper proposes reporting R^2 as a standard metric alongside topic models, as is typically done with OLS.

[6]Note, however, that the R^2 for the $K = 50$ (correctly-specified) LDA model is about 0.6. This is approximately the same value obtained for R^2 when using the population parameters.

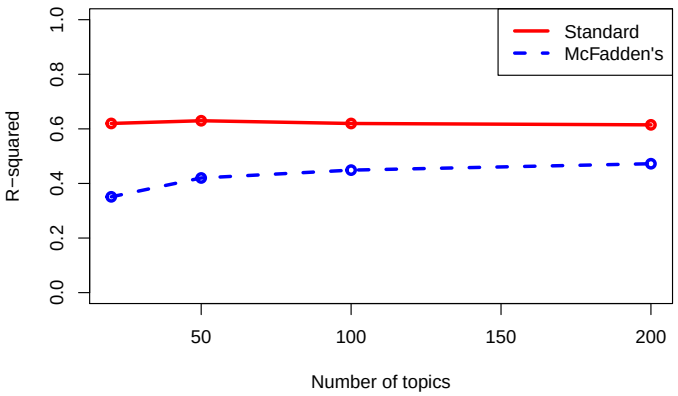

Fig. 2. Varying the number of topics on a simulated corpus

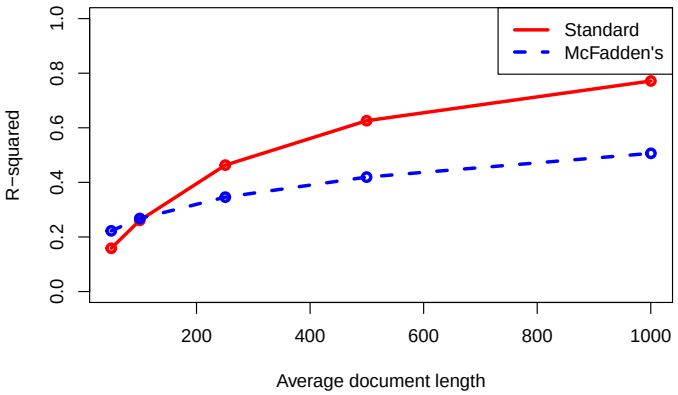

Fig. 3. Varying average document length on a simulated corpus

Acknowledgment

The authors would like to thank...

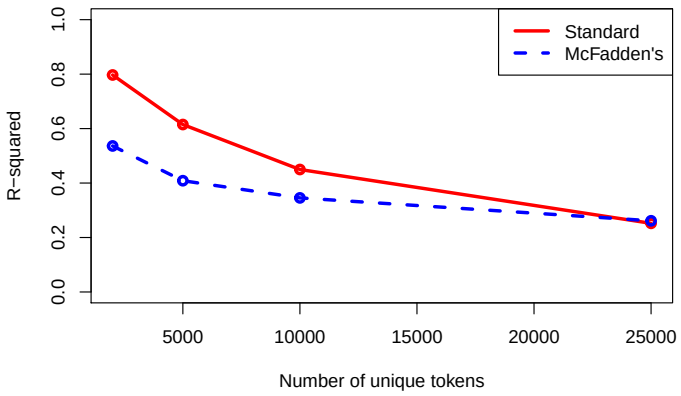

Fig. 4. Varying vocabulary size on a simulated corpus

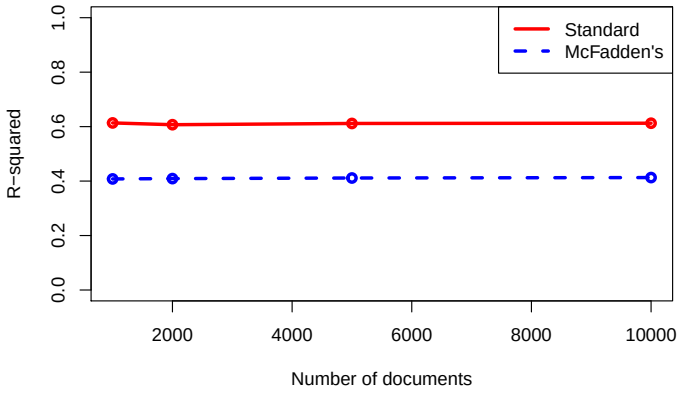

Fig. 5. Varying number of documents on a simulated corpus

References

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." Journal of Machine Learning Research 3 (Jan): 993–1022.

Buntine, Wray. 2009. "Estimating Likelihoods for Topic Models." In Asian Conference on Machine Learning, 51–64. Springer.

Girolami, Mark, and Ata Kabán. 2003. "On an Equivalence Between Plsi and Lda." In SIGIR, 3:433–34.