# $R^2$ FOR TOPIC MODELS

THOMAS W. JONES

Abstract. This document proposes a new (old) metric for evaluating goodness of fit in topic models, the coefficient of determination, or $R^2$.

## 1. Introduction

## 2. Probabilistic Topic Models

Probabilistic topic models are a family of stochastic models for estimating abstract "topics" in a set of documents. Many topic models have been developed and provide a flexible family of topic models. Some include frequently available metadata about documents, such as the time of the publication (in Dynamic Topic Models), or the author and location of the publication. Most topic models are Bayesian, though probabilistic latent semantic analysis (pLSA) is frequentist.[1] All probabilistic topic models share common features. Without loss of generality, all topic models model the document-generating process as a mixture of categorical distributions.[2] The goal of topic modeling is to estimate the parameters of these distributions. The most basic kind of topic model is parameterized as follows:

[1] pLSA is sometimes called probabilistic latent semantic indexing (pLSI).

[2] For estimation purposes, these are actually multinomial distributions with $N_d$ being the number of words in each document. The terms "multinomial" and "categorical" are often used interchangeably in the topic modeling literature.

$$\begin{aligned} Z_d &\sim Categorical_K(\theta_d) && d \in \{1, \dots, D\} \\ V_k &\sim Categorical_V(\phi_k) && k \in \{1, \dots, K\} \end{aligned}$$

where $Z_d$ represents topics over documents, $V_k$ represents words over topics, $K$ is the number of latent factors or "topics", $V$ is the number terms in the model and indexed by $v \in \{1, \dots, V\}$, and $N_d$ is the number of terms in the $d^{th}$ document.

A topic model's estimates are generally represented in two matrices., $\Theta$ and $\Phi$. The $d$-th row of $\Theta$ is $\theta_d$, whose $k$-th entry is the probability of topic $k$ conditional on document $d$. The $k$-th row of $\Phi$ is $\phi_k$, whose $v$-th entry is the probability of word $v$ conditional on topic $k$. The document term matrix $Y$ can be thought of as the result of repeated sampling from $\Theta$ and $\Phi$. The dot product of $\Phi$ and $\Theta$ is the expected value of the document term matrix: $E(Y) = \Theta \cdot \Phi^T$.

## 3. Goodness-of-Fit For Topic Models

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."[3] Common measures of fit include likelihood functions (or log likelihoods), information criterion such

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

as AIC or BIC, an area under a curve (AUC), and the coefficient of determination, or $R^2$. Goodness of fit measures may correct for model overfitting, such as in the case of AIC or BIC. "Perplexity" is a commonly-used measure in models of language. To compute perplexity, a document is held out of the fitting process. Half of the words in each document are used to estimate the most likely set of topic proportions; the model's goodness of fit is calculated on the held-out half of each document.

Goodness of fit is not the only consideration in statistical modeling. The goal of many statistical models, and topic models particularly, is inference. For example, the primary goal of a model predicting cancer recurrence rates may be to infer which patient behaviors are to be avoided or encouraged. Good inferential models are not always the best predictors. Many accurate and robust predictive models are "black boxes", making inference difficult. Researchers may trade off some goodness of fit for interpretability. However, goodness of fit is a key metric in establishing trust in a model. If an easily-interpreted model does not fit the data well, it cannot be trusted for inference.[4]

[PARAGRAPH HERE ON EVALUATION METRICS FOR TOPIC MODELS]

Topic models are generative models of word frequency. A common misconception holds that topic modeling is an "unsupervised" method.[5] Gooness of fit measures require outcome data against which to compare a model's fitted values.

[4] The opposite is not necessarily true; interpretability is not necessary for accurate prediction.

[5] Unsupervised methods are used on data where outcomes are unavailable or unneeded, such as in clustering.

[analogy: topics as coefficients in OLS.] However, the document term matrix, $Y$, contains outcomes: the words containd in the documents. As explained in the previous section, the dot product of the topic model's extimates give the expected value of $Y$; or $E(Y) = \hat{(}Y)$. Then the observed values, $Y$, can be compared to the fitted values, $\hat{Y}$ to measure a topic model's goodness of fit. In fact, this is the principle used in calculating the log likelihood of a model. By extension, we can calculate an $R^2$ for topic models.

## 4. Review: $R^2$ for the General Case

The common definition of $R^2$ is a ratio of summed squared errors.

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

For a model, $f$, of outcome variable, $y$, where there are $n$ observations, $R^2$ is derived as follows:

The mean of the data is $\bar{y} = E(y)$

The total sum of squares is $SS_{tot.} = \sum_{i-1}^{n} (y_i - \bar{y})^2$

The residual sum of squares is $SS_{resid.} = \sum_{i-1}^{n} (f_i - y_i)^2$

Finally, the coefficient of determination is $R^2 \equiv 1 - \frac{SS_{resid.}}{SS_{tot.}}$

$R^2$, as defined above, is bound between 0 and 1 when $f$ is a linear model such as ordinary least squares (OLS). If $f$ is not linear, then negative values of $R^2$ are possible. In this case, $SS_{resid.}$ may be larger than $SS_{tot.}$. However, when $f$ perfectly fit the data, $f_i = y_i \forall i$ and $SS_{resid} = 0$, making $R^2$ is equal to one.

In some cases, $SS_{tot.}$ can be partitioned into two parts: $SS_{tot.} = SS_{resid.} + SS_{model}$. $SS_{model}$ is the model sum of squares: $SS_{model} = \sum_{i-1}^{n} (f_i - \bar{y})^2$. When this relation holds, then $R^2$ is interpreted as the proportion of variance of $y$ explained by $f$. This is derived below.[6] (This relationship does not hold for topic modeling, however.)

[6] TOMMY, YOU NEED A REFERENCE FOR THIS AS THIS DERIVATION DOES NOT PROVE THAT $E(f) = y$. Also, does this interpretation hold in the Bayesian case? Must investigate.

$$\begin{aligned}
V(f) &= E((f - E(f))^2) \\
&= E((f - E(y))^2) \\
&= E((f - \bar{y})) \\
&= \frac{1}{n}\sum_{i-1}^{n}(f_i - \bar{y})^2 \\
&= \frac{1}{n}SS_{model} \\
V(y) &= E((y - E(y))^2) \\
&= E((y - \bar{y})^2) \\
&= \frac{1}{n}\sum_{i-1}^{n}(y_i - \bar{y})^2 \\
&= \frac{1}{n}SS_{tot.} \\
\frac{V(f)}{V(y)} &= \frac{SS_{model}/n}{SS_{tot.}/n} \\
&= \frac{SS_{model}}{SS_{tot.}} \\
&= 1 - \frac{SS_{resid.}}{SS_{tot.}}7 \\
&\equiv R^2
\end{aligned}$$

## 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 $p, q \in \mathbb{R}_m$

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

where $d(p,q)$ denotes the euclidean distance between $p$ and $q$. $R^2$ 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 $d(y_i, \bar{y}) = \sqrt{(y_i - \bar{y})^2}$.[8] In the multidimensional case where $y_i, f_i \in \mathbb{R}_m; m > 1$, then $\bar{y} \in \mathbb{R}_m$ represents the point at the center of $y$ in $m$ space.

We can rewrite $R^2$ using the relationships above.

$$\begin{aligned} SS_{tot.} &= \sum_{i=1}^{n} d(y_i, \bar{y})^2 \\ SS_{resid.} &= \sum_{i=1}^{n} d(y_i, f_i)^2 \\ \Rightarrow R^2 &= 1 - \frac{\sum_{i=1}^{n} d(y_i, f_i)^2}{\sum_{i=1}^{n} d(y_i, \bar{y})^2} \end{aligned}$$

Figure [1] depicts

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,

[8] 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 $y_i, f_i \in \mathbb{R}_m; m > 1$.

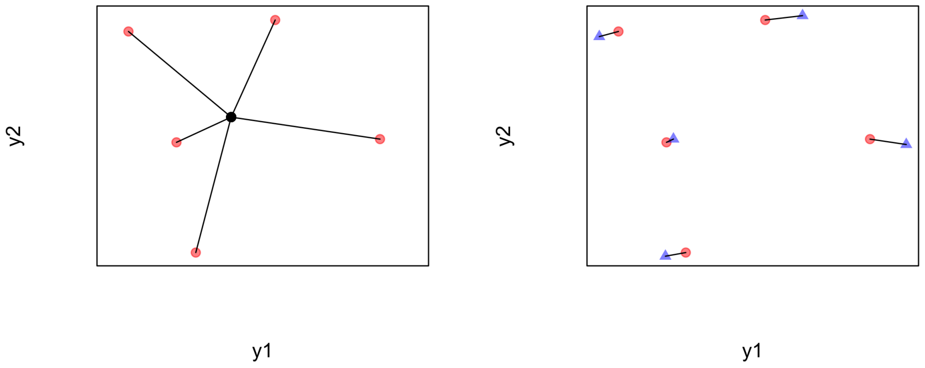

FIGURE 1. The figure on the left depicts... the figure on the right depicts...

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.

## 6. $R^2$ For Topic Models

An $R^2$ for topic models follows from the geometric interpretation of $R^2$. [ paragraph follows ]

$$explicitderivationhere$$

The $R^2$ for topic models is still the coefficient of determination. Several psuedo cofficients of determination have been made for various types of data and model. [Explain one or two.] However, the $R^2$ for topic models proposed in this paper follows from the traditional definition of $R^2$.

## 7. Advantages of Using $R^2$ for Topic Models

Using $R^2$ eases communication of goodness of fit in topic modeling. [It's easy to interpret. People are used to it. Being maximized at 1 (and generally not more than zero), facillitates comparing models across corpra. This lets us develop rules of thumb for "good" models in different contexts.]

## 8. Use of $R^2$ on Simulated Data

Latent Dirichlet allocation (LDA), possibly the most popular topic model, places Dirichlet priors on $\theta_d, \forall d$ and $\phi_k, \forall k$:

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

## 9. Use of $R^2$ on NIH Award Abstracts

## 10. Conclusion

## 11. Appendix

## References