Measuring Confidence in Temporal Topic Models with Posterior Predictive Checks

David Mimno
Department of Computer Science
Princeton University Princeton, NJ 08540
mimno@cs.princeton.edu

David Blei
Department of Computer Science
Princeton University Princeton, NJ 08540
blei@cs.princeton.edu

Abstract

Large text collections are useful in social science research, but building reliable predictive models is difficult. Researchers must either deal directly with sparse, noisy, high dimensional language data or use latent variable models to infer more tractable lower dimensional patterns. For conclusions based on latent variable models to be reliable, however, it is necessary to measure the degree to which the model's assumptions are met and to verify that the inferred hidden structure adequately captures the important variation in the data. In this paper we present one such methodology for evaluating statistical topic models using posterior predictive checks. As an example, we consider the content of 19th century speeches from the House of Commons.

1 Introduction

Text analysis has begun to provide a useful complement to traditional methods in social science research. Document collections are often available in historical and cultural contexts that would be difficult or impossible to address with surveys. Text analysis also raises many new problems: language data is high dimensional, highly variable, and sparse. As a result, reliably distinguishing "signal" — changes in language that result from real changes in the social and intellectual environment — from "noise" — random variation in word choice — is a difficult problem. Latent variable models such as statistical topic models provide an attractive alternative to working directly with word count data [1]. These models assert that, conditioned on some hidden structure (for example topic distributions for documents), words are selected independently. If this claim is true, then the much lower-dimensional latent representation can be substituted for the actual observed words without any consequences for subsequent analysis. When considering inferences drawn from such models we would like to measure the degree to which these assumptions are violated. In this paper we explore methods for checking the validity of latent topic models. As a running example, we use speeches from the British House of Commons from the period 1830–1891.[1]

A common question in text analysis is whether there is an association between observable conditions at particular times and the text of documents written during those periods in a particular collection. In our example, explanatory variables might include the 16 parliaments that were elected during this time period, the 10 prime ministers, and five political parties that were in power. For the purposes of this work we only consider one such variable, time of publication. Manually coding documents according to a controlled vocabulary would be expensive, time-consuming, and prone to annotator bias. Directly analyzing the relationship between an explanatory variable X and a word variable W is also challenging: there are many words and almost all of them are rare, resulting in many poorly estimated parameters. Rather than analyzing the association between words and parliaments or PMs directly, we might use a topic model as a pre-processing step to reduce the dimensionality of the

[1] Available from http://hansard-archive.parliament.uk

data. The topic model assumes that the choice of observed words is independent, when conditioned on the identities of their topics. We can then assess the association of the random variables X to the topics Z, with the assumption that X is then conditionally independent of W given Z.

2 Measuring model assumptions

Moving from the tens of thousands of dimensions typical of vector-space representations of text documents to the few hundred dimensions typical of topic models necessarily throws out information. We would like to analyze text by substituting topics for the individual words, but we need to be able to estimate how much remaining dependence exists between explanatory variables and words given topics. The distribution of words varies over time. Ideally, we would like all of this variation to be explained by changes in topic concentration over time, rather than by changes in the probability of words given topics. This criterion suggests a method for checking the model: for each topic, measure the discrepancy between the empirical word distribution at each time step and the overall distribution over words for the topic.

We consider two measures of this association. Let $N(w,t,k)$ be the number of tokens of type w in topic k at time t, with $N(w,k) = \sum_t N(w,t,k)$, $N(t,k) = \sum_w N(w,t,k)$, and $N(k) = \sum_{w,t} N(w,t,k)$. The mutual information of w and t is

$$MI(w,t \mid k) = \sum_w \sum_t P(w,t \mid k) \log \frac{P(w \mid t,k)P(t \mid k)}{P(w \mid k)P(t \mid k)} \quad (1)$$

$$= \sum_w \sum_t \frac{N(w,t,k)}{N(k)} \log \frac{N(w,t,k)N(k)}{N(t,k)N(w,k)} \quad (2)$$

This metric measures the divergence between the joint distribution over word and time and the product of the marginal distributions. A mutual information of 0 implies independence. Another measure of divergence from an independence assumption for multinomial observations is the Q score [3]:

$$Q(w,t \mid k) = \sum_t \sum_w \frac{\left(N(w,t,k) - N(t,k)\frac{N(w,k)}{N(k)}\right)^2}{N(t,k)\frac{N(w,k)}{N(k)}} \quad (3)$$

where $N(t,k)\frac{N(w,k)}{N(k)}$ is the expected number of tokens of word type w in a multinomial observation of length $N(t,k)$ with estimated probability $\hat{p}_{w|k} = \frac{N(w,k)}{N(k)}$. Again, smaller values are more consistent with the assumption of conditional independence of words given topics.

3 Posterior predictive checks

Evaluating these functions, however, is not sufficient, as they are to some extent a function of the distribution of the topic over time. In the extreme case, a topic that only occurs in one time step t will have 0 mutual information and 0 Q-score because $P(w,t \mid k) = P(w \mid k)$ and $P(t \mid k)$ is deterministic, while another topic with the same number of tokens evenly distributed over all time steps is unlikely to have exactly the same empirical distribution over words for each time step.

In order to interpret such values, we must compare them to some reference distribution, so as to determine whether they are typical or out of the ordinary. Traditional frequentist p-values are one example of such contextualization, where the reference distribution is usually some asymptotic distribution such as a Student t or χ^2. Such asymptotic results are difficult to formulate for models with latent variables. Rather, we will use a *posterior predictive check*, a Bayesian method for assessing model fitness [5]. Consider an observed variable w, a probabilistic model $P(w|\Theta)$ where Θ is a set of fixed hyperparameters, and some function of interest $f(w)$, which we will call a discrepancy function. In a PPC, $f(w)$ is compared to a reference distribution derived by repeatedly sampling new values w^{rep} from the posterior distribution $P(w^{rep} \mid \Theta, w)$, that is, a distribution trained on the observed data w, holding certain conditions such as the sample size fixed. We can then evaluate $P(f(w^{rep}) > f(w) \mid w)$ by counting the number of values of w^{rep} that result in

greater values of the function of interest than the value for the observed data. As originally formulated, PPCs are challenging in Bayesian hierarchical models with latent variables (that is, where $P(w|\Theta) = \sum_z P(w|z)P(z|\Theta)$) because sampling from the posterior distribution over observed variables involves marginalizing over the hidden variables. Gelman, Meng, and Stern [2] introduce the method of realized discrepancies for PPCs, in which MCMC methods are used to draw samples of the hidden variables (that is, from $P(z|\Theta, w)$), which can then be used to draw replications of the observed data, from $P(w^{rep}|z, w)$.

In this work we perform a posterior predictive check for each topic individually, resulting in a vector-valued discrepancy function rather than a scalar function. Let $\boldsymbol{w}$ be the words assigned to some topic k at a particular state of a Gibbs sampler. Given $N(w, k)$, we can create a replicated vector $\boldsymbol{w}^{rep}$ of the same length by repeatedly sampling w_i^{rep} with probability proportional to $N(w, k)$, such that $\sum_i N(w_i^{rep}, t, k) = N(t, k)$, thus holding the topic counts fixed per time step. We can then reevaluate the mutual information and Q-score for the replication.

4 Results

As an example, we consider a corpus consisting of 543,112 speeches from the British House of Commons from 1830–1891. The collection consists of 55 million words from a vocabulary of size 50,000 after removal of most frequent and least frequent word types, divided into 305 volumes, each comprising about three weeks, with between 600 and 4000 speeches per session. We trained topic models with the number of topics $K \in \{100, 150, 200, 250, 300, 1000, 1500\}$ with 2000 iterations of Gibbs sampling, using the Mallet topic modeling package [4]. Hyperparameters were optimized every 20 iterations. For the following experiments we used the 305 volumes as the time indicators t in Equations 1 and 3. Models had no access to volume information at training time.

Before exploring the variability of word choice within topics, we measure the overall variability of the collection by calculating Equations 1 and 3 on the entire corpus, that is $MI(w, t)$ and $Q(w, t)$. We refer to this as the "words" model, which is equivalent to a topic model with $K = 1$. We can compare these measurements to the same calculation after substituting topic assignments for words: that is, the association between topics and time $MI(k, t)$ and $Q(k, t)$. These numbers are not necessarily comparable across models, as the number of dimensions varies between 100 for the smallest topic model and 50,000 for the "words" model. We therefore draw 100 replications of the corpus (holding document lengths fixed) from the posterior of each model, that is, $p(w)$ for the "words" model and $p(k)$ for each topic model. Results are shown in Figure 1. Note that since our goal is to use these topic variables to make claims about temporal patterns, a deviation from the null model that topics are i.i.d. multinomial is desirable. Values for the "words" model are comparable to values for the topic models. Association between topics and time increases as the number of topics increases. The mean replicated values are different: they are small for the topic models, while the "words" model is only three to four times larger than the expected value. The large difference in expected values between word and topic models is due to the number and sparsity of dimensions.

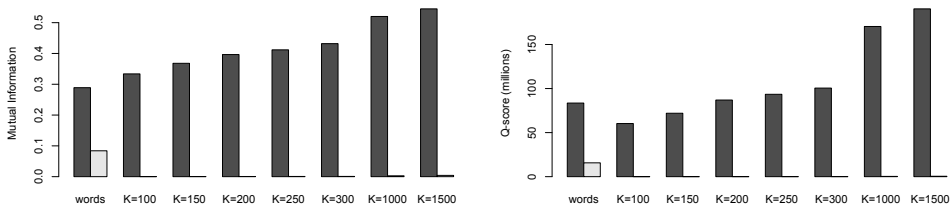

Figure 1: **Topics have strong association with time, relative to words alone.** For each model, the left bar is the observed value of the metric, and the right bar is the mean value of replications, which is very small for all topic models. Standard errors are not visible at this scale.

Figure 2 shows a comparison of the actual values of mutual information and Q-score alongside mean values of the same metrics averaged over 100 replications from the posterior distribution over words for each topic. In contrast to the previous section, deviation from the multinomial model indicates that there is additional pattern that will be hidden by replacing words with topics, and is

therefore undesirable. The figure shows topics from three models, with $K \in \{150, 300, 1000\}$. Each point represents a topic. Replicated topic-word distributions have consistently smaller values than the observed values of both metrics, indicating that words are more "concentrated" in time than the model expects. In order to provide insight into the nature of temporal variation, we can group the terms in the summation in Equation 1 by word and rank the words by their contribution to the discrepancy function. For example, a topic with the most probable words *ships, vessels, admiralty, iron, ship, navy* has as its most "mismatching" words *iron (0.0065), turret (0.0038), clads (0.0036), wooden (0.0036)*, consistent with changes in naval technology during the Victorian era (that is, wooden ships to "iron clads").

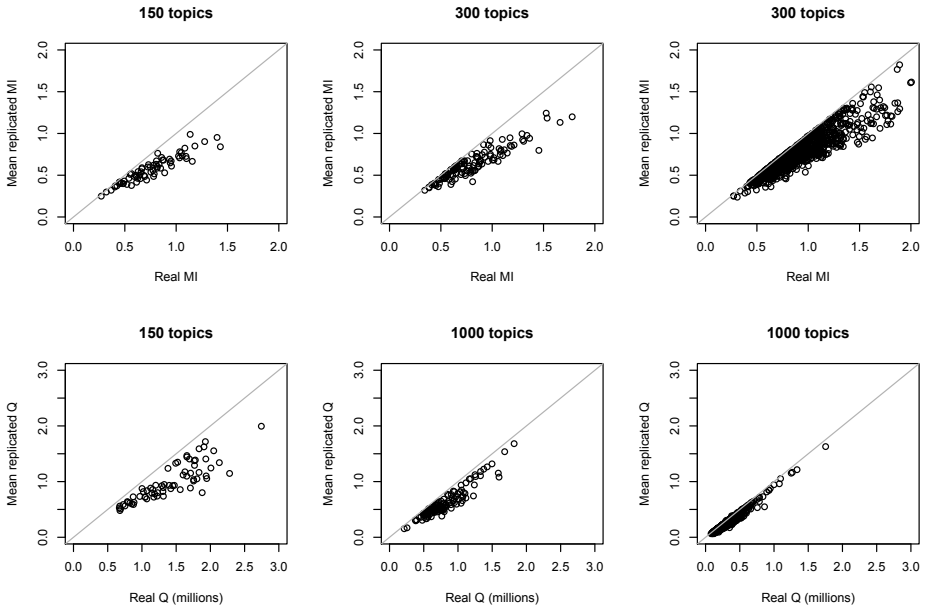

Figure 2: **Observed word-time associations are slightly greater than the mean replicated values.** Actual topic metrics (x-axis) to the mean value of the same metric from replications drawn from the posterior for each topic (y-axis), with a line showing $x = y$. At $K = 300$ R is 0.93 (MI) and 0.97 (Q), but variability is high: $MI^{rep}(w, t|k)/MI^{obs}(w, t|k)$ has mean 0.81 ± 0.10

In addition to the mean, we can also consider the variance of the replicated values, which tend to be very consistent between replications. The average number of standard deviations, as estimated from the posterior distributions for models with $K \in \{150, 200, 250, 300, 1000, 1500\}$ is $135, 121, 99, 80, 35, 26$ for MI and $33, 26, 21, 19, 8, 6$ for Q, respectively. The mismatch between observed values and expected values under the posterior distributions is thus highly unlikely to be the result of random chance. The mismatch seems to decrease as K increases.

Our goal is to measure the confidence with which we can replace words with topic indicators in a temporal analysis of a text collection. Although we trained models without any access to time information, the resulting topic variables show an increased association with time relative to words alone, possibly due to their ability to cluster rare words and disambiguate frequent words. Using PPCs, we have also determined that (1) there is a consistent, statistically significant divergence between the patterns of words actually assigned to topics and random samples from the same topics' posterior distributions holding the number of words per time step fixed, and (2) there is substantial variation in the magnitude of this mismatch from one topic to another. This mismatch suggests that although topics are useful in identifying temporal trends, results should be accompanied by topic quality metrics. The methods used in this work suggest ways for calculating such metrics and further exploring the reasons for observed variability within topics.

Acknowledgments

Arthur Spurling and Andy Eggers suggested the use of the Hansards corpus.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[2] A. Gelman, X. Meng, and H. Stern. posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

[3] G. Kanji. *100 Statistical Tests*. SAGE, 2006.

[4] A. K. McCallum. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[5] D. B. Rubin. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6:377–401, 1981.