Bayesian Checking for Topic Models

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

Real document collections do not fit the independence assumptions asserted by most statistical topic models, but how badly do they violate them? We present a Bayesian method for measuring how well a topic model fits a corpus. Our approach is based on posterior predictive checking, a method for diagnosing Bayesian models in user-defined ways. Our method can identify where a topic model fits the data, where it falls short, and in which directions it might be improved.

1 Introduction

Probabilistic topic models are a suite of machine learning algorithms that decompose a corpus into a set of topics and represent each document with a subset of those topics. The inferred topics often correspond with the underlying themes of the analyzed collection, and the topic modeling algorithm organizes the documents according to those themes.

Most topic models are evaluated by their predictive performance on held out data. The idea is that topic models are fit to maximize the likelihood (or posterior probability) of a collection of documents, and so a good model is one that assigns high likelihood to a held out set (Blei et al., 2003; Wallach et al., 2009).

But this evaluation is not in line with how topic models are frequently used. Topic models seem to capture the underlying themes of a collection—indeed the monicker "topic model" is retrospective—and so we expect that these themes are useful for exploring, summarizing, and learning about its documents (Mimno and McCallum, 2007; Chang et al., 2009). In such exploratory data analysis, however, we are not concerned with the fit to held out data.

In this paper, we develop and study new methods for evaluating topic models. Our methods are based on *posterior predictive checking*, which is a model diagnosis technique from Bayesian statistics (Rubin, 1984; Gelman et al., 1996). The goal of a posterior predictive check (PPC) is to assess the validity of a Bayesian model without requiring a specific alternative model. Given data, we first compute a posterior distribution over the latent variables. Then, we estimate the probability of the observed data under the data-generating distribution that is induced by the posterior (the "posterior predictive distribution"). A data set that is unlikely calls the model into question, and consequently the posterior. PPCs can show where the model fits and doesn't fit the observations. They can help identify the parts of the posterior that are worth exploring.

The key to a posterior predictive check is the *discrepancy function*. This is a function of the data that measures a property of the model which is important to capture. While the model is often chosen for computational reasons, the discrepancy function might capture aspects of the data that are desirable but difficult to model. In this work, we will design a discrepancy function to measure an independence assumption that is implicit in the modeling assumptions but is not enforced in the posterior. We will embed this function in a posterior predictive check and use it to evaluate and visualize topic models in new ways.

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 227–237,
Edinburgh, Scotland, UK, July 27–31, 2011.

Specifically, we develop discrepancy functions for latent Dirichlet allocation (the simplest topic model) that measure how well its statistical assumptions about the topics are matched in the observed corpus and inferred topics. LDA assumes that each observed word in a corpus is assigned to a topic, and that the words assigned to the same topic are drawn independently from the same multinomial distribution (Blei et al., 2003). For each topic, we measure the whether this assumption holds by computing the mutual information between the words assigned to that topic and which document each word appeared in. If the assumptions hold, these two variables should be independent: low mutual information indicates that the assumptions hold; high mutual information indicates a mismatch to the modeling assumptions.

We embed this discrepancy in a PPC and study it in several ways. First, we focus on topics that model their observations well; this helps separate interpretable topics from noisy topics (and "boilerplate" topics, which exhibit too little noise). Second, we focus on individual terms within topics; this helps display a model applied to a corpus, and understand which terms are modeled well. Third, we replace the document identity with an external variable that might plausibly be incorporated into the model (such as time stamp or author). This helps point the modeler towards the most promising among more complicated models, or save the effort in fitting one. Finally, we validate this strategy by simulating data from a topic model, and assessing whether the PPC "accepts" the resulting data.

2 Probabilistic Topic Modeling

Probabilistic topic models are statistical models of text that assume that a small number of distributions over words, called "topics," are used to generate the observed documents. One of the simplest topic models is latent Dirichlet allocation (LDA) (Blei et al., 2003). In LDA, a set of K topics describes a corpus; each document exhibits the topics with different proportions. The words are assumed exchangeable within each document; the documents are assumed exchangeable within the corpus.

More formally, let $\phi_1, \ldots, \phi_K$ be K topics, each of which is a distribution over a fixed vocabulary. For each document, LDA assumes the following generative process

1. Choose topic proportions $\theta_d \sim \text{Dirichlet}(\alpha)$.
2. For each word
   (a) Choose topic assignment $z_{d,n} \sim \theta$.
   (b) Choose word $w_{d,n} \sim \phi_{z_{d,n}}$.

This process articulates the statistical assumptions behind LDA: Each document is endowed with its own set of topic proportions θ_d, but the same set of topics $\phi_{1:K}$ governs the whole collection.

Notice that the probability of a word is independent of its document θ_d given its topic assignment $z_{d,n}$ (i.e., $w_{d,n} \perp\!\!\!\perp \theta_d \,|\, z_{d,n}$). Two documents might have different overall probabilities of containing a word from the "vegetables" topic; however, all the words in the collection (regardless of their documents) drawn from that topic will be drawn from the same multinomial distribution.

The central computational problem for LDA is posterior inference. Given a collection of documents, the problem is to compute the conditional distribution of the hidden variables—the topics ϕ_k, topic proportions θ_d, and topic assignments $z_{d,n}$. Researchers have developed many algorithms for approximating this posterior, including sampling methods (Griffiths and Steyvers, 2004) (used in this paper), variational methods (Blei et al., 2003), distributed variants (Asuncion et al., 2008), and online algorithms (Hoffman et al., 2010).

3 Checking Topic Models

Once approximated, the posterior distribution is used for the task at hand. Topic models have been applied to many tasks, such as classification, prediction, collaborative filtering, and others. We focus on using them as an exploratory tool, where we assume that the topic model posterior provides a good decomposition of the corpus and that the topics provide good summaries of the corpus contents.

But what is meant by "good"? To answer this question, we turn to Bayesian model checking (Rubin, 1981; Gelman et al., 1996). The goal of Bayesian model checking is to assess whether the observed data matches the modeling assumptions in the directions that are important to the analysis. The

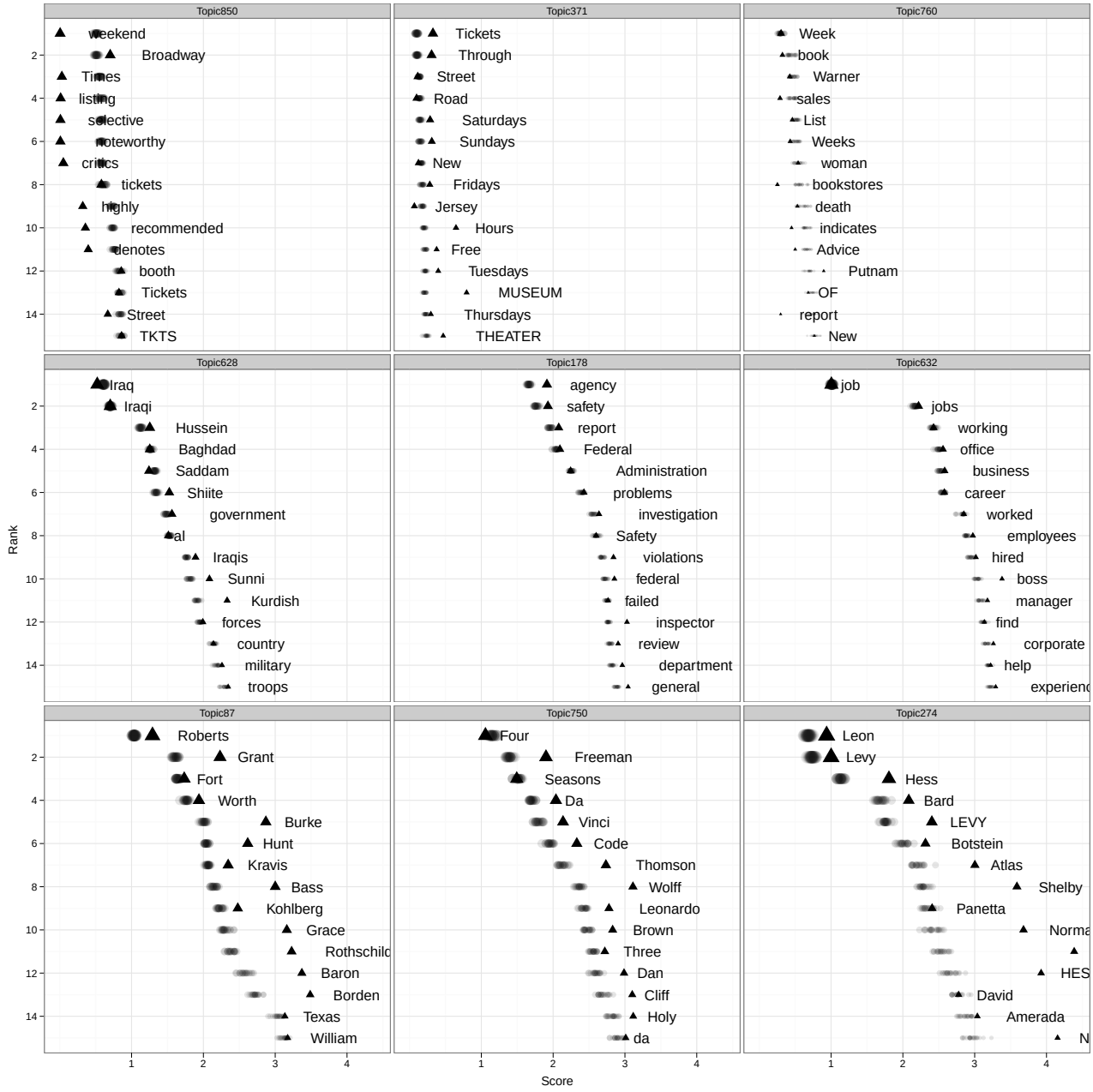

Figure 1: **Visualization of variability within topics.** Nine randomly selected topics from the New York Times with low (top row), medium (middle row) and high (bottom row) mutual information between words and documents. The y-axis shows term rank within the topic, with size proportional to log probability. The x-axis represents divergence from the multinomial assumption for each word: terms that are uniformly distributed across documents are towards the left, while more specialized terms are to the right. Triangles represent real values, circles represent 20 replications of this same plot from the posterior of the model.

intuition is that only when satisfied with the model should the modeler use the posterior to learn about her data. In complicated Bayesian models, such as topic models, Bayesian model checking can point to the parts of the posterior that better fit the observed data set and are more likely to suggest something meaningful about it.

In particular, we will develop posterior predictive checks (PPC) for topic models. In a PPC, we specify a *discrepancy function*, which is a function of the data that measures an important property that we want the model to capture. We then assess whether the observed value of the function is similar to values of the function drawn from the posterior, through the distribution of the data that it induces. (This distribution of the data is called the "posterior predictive distribution.")

An innovation in PPCs is the *realized discrepancy function* (Gelman et al., 1996), which is a function of the data and any hidden variables that are in the model. Realized discrepancies induce a traditional discrepancy by marginalizing out the hidden variables. But they can also be used to evaluate assumptions about *latent* variables in the posterior, especially when combined with techniques like MCMC sampling that provide realizations of them. In topic models, as we will see below, we use a realized discrepancy to factor the observations and to check specific components of the model that are discovered by the posterior.

3.1 A realized discrepancy for LDA

Returning to LDA, we design a discrepancy function that checks the independence assumption of words given their topic assignments. As we mentioned above, given the topic assignment z the word w should be independent of its document θ. Consider a decomposition of a corpus from LDA, which assigns every observed word $w_{d,n}$ to a topic $z_{d,n}$. Now restrict attention to all the words assigned to the kth topic and form two random variables: W are the words assigned to the topic and D are the document indices of the words assigned to that topic. If the LDA assumptions hold then knowing W gives no information about D because the words are drawn independently from the topic.

We measure this independence with the mutual information between W and D:[1]

$$\begin{aligned} &MI(W, D \mid k) \\ &= \sum_w \sum_d P(w, d \mid k) \log \frac{P(w \mid d, k) P(d \mid k)}{P(w \mid k) P(d \mid k)} \\ &= \sum_w \sum_d \frac{N(w, d, k)}{N(k)} \log \frac{N(w, d, k) N(k)}{N(d, k) N(w, k)}. \end{aligned} \quad (1)$$

Where $N(w, d, k)$ is the number of tokens of type w in topic k in document d, with $N(w, k) = \sum_d N(w, d, k)$, $N(d, k) = \sum_w N(w, d, k)$, and $N(k) = \sum_{w,d} N(w, d, k)$. This function measures the divergence between the joint distribution over word and document index and the product of the marginal distributions. In the limit of infinite samples, independent random variables have mutual information of zero, but we expect finite samples to have non-zero values even for truly independent variables. Notice that this is a realized discrepancy; it depends on the latent assignments of observed words to topics.

Eq. 1 is defined as a sum over a set of documents and a set of words. We can rearrange this summation as a weighted sum of the *instantaneous mutual information* between words and documents:

$$IMI(w, D \mid k) = H(D|k) - H(D \mid W = w, k). \quad (2)$$

This quantity can be understood by considering the per-topic distribution of document labels, $p(d|k)$. This distribution is formed by normalizing the counts of how many words assigned to topic k appeared in each document. The first term of Eq. 2 is the entropy—some topics are evenly distributed across many documents (high entropy); others are concentrated in fewer documents (low entropy).

The second term conditions this distribution on a particular word type w by normalizing the per-document number of times w appeared in each document (in topic k). If this distribution is close to $p(d|k)$ then $H(D|W = w, k)$ will be close to $H(D|k)$ and $IMI(w, D|k)$ will be low. If, on the other hand, word w occurs many times in only a few documents, it will have lower entropy over docu-

[1] There are other choices of discrepancies, such as word-word point-wise mutual information scores (Newman et al., 2010).

ments than the overall distribution over documents for the topic and $IMI(w, D|k)$ will be high.

We illustrate this discrepancy in Figure 1, which shows nine topics trained from the *New York Times*.[2] Each row contains randomly selected topics from low, middle, and high ranges of MI, respectively. Each triangle represents a word. Its place on the y-axis is its rank in the topic. Its place on the x-axis is its $IMI(w|k)$, with more uniformly distributed words (low IMI) to the left and more specific words (high IMI) to the right. (For now, ignore the other points in this figure.) IMI varies between topics, but tends to increase with rank as less frequent words appear in fewer documents.

The discrepancy captures different kinds of structure in the topics. The top left topic represents formulaic language, language that occurs verbatim in many documents. In particular, it models the boilerplate text "Here is a selective listing by critics of The Times of new or noteworthy..." Identifying repeated phrases is a common phenomenon in topic models. Most words show lower than expected IMI, indicating that word use in this topic is less variable than data drawn from a multinomial distribution. The middle-left topic is an example of a good topic, according to this discrepancy, which is related to Iraqi politics. The bottom-left topic is an example of the opposite extreme from the top-left. It shows a loosely connected series of proper names with no overall theme.

3.2 Posterior Predictive Checks for LDA

Intuitively, the middle row of topics in Figure 1 are the sort of topics we look for in a model, while the top and bottom rows contain topics that are less useful. Using a PPC, we can formally measure the difference between these topics. For each of the real topics in Figure 1 we regenerated the same figure 20 times. We sampled new words for every token from the posterior distribution of the topic, and recalculated the rank and IMI for each word. These "shadow" figures are shown as gray circles. The density of those circles creates a reference distribution indicating the expected IMI values at each rank under the multinomial assumption.

[2]Details about the corpus and model fitting are in Section 4.2. Similar figures for two other corpora are in the supplement.

By themselves, IMI scores give an indication of the distribution of a word between documents within a topic: small numbers are better, large numbers indicate greater discrepancy. These scores, however, are based on the specific allocation of words to topics. For example, lower-ranked, less frequent words within a topic tend to have higher IMI scores than higher-ranked, more frequent words. This difference may be due to greater violation of multinomial assumptions, but may also simply be due to smaller sample sizes, as the entropy $H(D|W = w, k)$ is estimated from fewer tokens. The reference distributions help distinguish between these two cases.

In more detail, we generate replications of the data by considering a Gibbs sampling state. This state assigns each observed word to a topic. We first record the number of instances of each term assigned to each topic, $N(w|k)$. Then for each word $w_{d,n}$ in the corpus, we sample a new observed word $w_{d,n}^{rep}$ where $P(w) \propto N(w|z_{d,n})$. (We did not use smoothing parameters.) Finally, we recalculate the mutual information and instantenous mutual information for each topic.

In the top-left topic, most of the words have much lower IMI than the word at the same rank in replications, indicating lower than expected variability. The exception is the word *Broadway*, which is more variable than expected. In the middle-left topic, IMI for the words *Iraqi* and *Baghdad* occur within the expected range. These words fit the multinomial assumption: any word assigned to this topic is equally likely to be *Iraqi*. Values for the words *Shiite*, *Sunni*, and *Kurdish* are more specific to particular documents than we expect under the model. In the bottom-left topic, almost all words occur with greater variability than expected. This topic combines many terms with only coincidental similarity, such as Mets pitcher Grant Roberts and the firm Kohlberg Kravis Roberts.

Turning to an analysis of the full mutual information, Figure 2 shows the three left-hand topics from Figure 1: *Weekend*, *Iraq*, and *Roberts*. The histogram represents MI scores for 100 replications of the topic, rescaled to have mean zero and unit variance. The observed value, also rescaled, and the mean replicated value (set to zero) are shown with vertical lines. The formulaic *Weekend* topic has significantly lower than expected MI. The *Iraq*

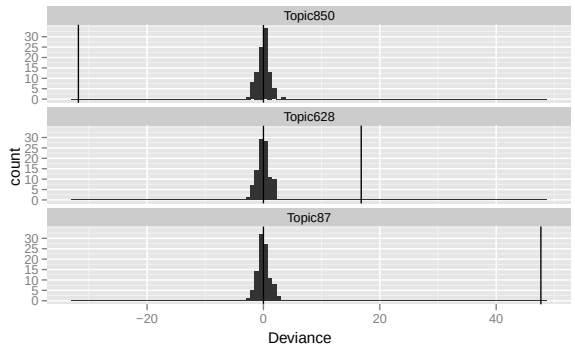

Figure 2: **News: Observed topic scores (vertical lines) relative to replicated scores, rescaled so that replications have zero mean and unit variance.** The *Weekend* topic (top) has lower than expected MI. The *Iraq* (middle) and *Roberts* (bottom) topics both have MI greater than expected.

and *Roberts* topics have significantly greater than expected MI.

For most topics the actual discrepancy is outside the range of any replicated discrepancies. In their original formulation, PPCs prescribe computing a tail probability of a replicated discrepancy being greater than (or less than) the observed discrepancy under the posterior predictive distribution. For example if an observed value is greater than 70 of 100 replicated values, we report a PPC p-value of 0.7.

When the observed value is far outside the range of any replicated values, as in Figure 2, that tail probability will be degenerate at 0 or 1. So, we report instead a *deviance* value, an alternative way of comparing an observed value to a reference distribution. We compute the distribution of the replicated discrepancies and compute its standard deviation. We then compute how many standard deviations the observed discrepancy is from the mean of the replicated discrepancies.

This score allows us to compare topics. The observed value for the *Weekend* topic is 31.8 standard deviations below the mean replicated value, and thus has deviance of -31.8, which is lower than expected. The *Iraq* topic has deviance of 16.8 and the *Roberts* topic has deviance of 47.7. This matches our intuition that the former topic is more useful than the latter.

4 Searching for Systematic Deviations

We demonstrated that the mutual information discrepancy function can detect violations of multinomial assumptions, in which instances of a term in a given topic are not independently distributed among documents. One way to address this lack of fit is to encode document-level extra-multinomial variance ("burstiness") into the model using Dirichlet compound multinomial distributions (Doyle and Elkan, 2009). If there is no pattern to the deviations from multinomial word use across documents, this method is the best we can do.

In many corpora, however, there are systematic deviations that can be explained by additional variables. LDA is the simplest generative topic model, and researchers have developed many variants of LDA that account for a variety of variables that can be found or measured with a corpus. Examples include models that account for time (Blei and Lafferty, 2006), books (Mimno and McCallum, 2007), and aspect or perspective (Mei and Zhai, 2006; Lin et al., 2008; Paul et al., 2010). In this section, we show how we can use the mutual information discrepancy function of Equation 1 and PPCs to guide our choice in which topic model to fit.

Greater deviance implies that a particular grouping better explains the variation in word use within a topic. The discrepancy functions are large when words appear more than expected in some groups and less than expected in others. We know that the individual documents show significantly more variation than we expect from replications from the model's posterior distribution. If we combine documents randomly in a meaningless grouping, such deviance should decrease, as differences between documents are "smoothed out." If a grouping of documents shows equal or greater deviation, we can assume that that grouping is maintaining the underlying structure of the systematic deviation from the multinomial assumption, and that further modeling or visualization using that grouping might be useful.

4.1 PPCs for systematic discrepancy

The idea is that the words assigned to a topic should be independent of both document and any other variable that might be associated with the document. We simply replace the document index d with another

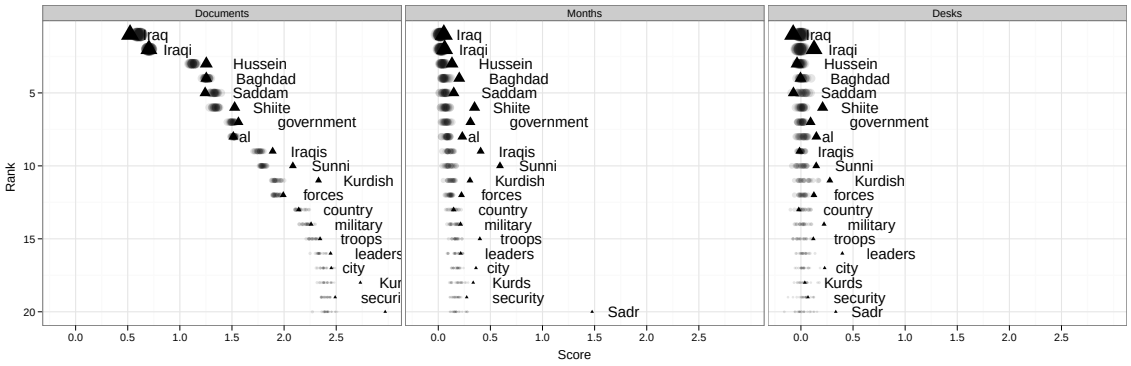

Figure 3: **Groupings decrease MI, but values are still larger than expected.** Three ways of grouping words in a topic from the *New York Times*. The word *leaders* varies more between desks than by time, while *Sadr* varies more by time than desk.

variable g in the discrepancy. For example, the New York Times articles are each associated with a particular news desk and also associated with a time stamp. If the topic modeling assumptions hold, the words are independent of both these variables. If we see a significant discrepancy relative to a grouping defined by a metadata feature, this systematic variability suggests that we might want to take that feature into account in the model.

Let $\mathcal{G}$ be a set of groups and let $\gamma \in \mathcal{G}^D$ be a grouping of D documents. Let $N(w, g, k) = \sum_d N(w, d, k) I_{\gamma_d = g}$, that is, the number of words of type w in topic k in documents in group g, and define the other count variables similarly. We can now substitute these group-specific counts for the document-specific counts in the discrepancy function in Eq. 1. Note that the previous discrepancy functions are equivalent to a trivial grouping, in which each document is the only member of its own group. In the following experiments we explore groupings by published volume, blog, preferred political candidate, and newspaper desk, and evaluate the effect of those groupings on the deviation between mean replicated values and observed values of those functions.

4.2 Case studies

We analyze three corpora, each with its own metadata: the New York Times Annotated Corpus (1987–2007)[3], the CMU 2008 political blog corpus (Eisenstein and Xing, 2010), and speeches from the British House of Commons from 1830–1891.[4] Descriptive statistics are presented in Table 1. The realization is represented by a single Gibbs sampling state after 1000 iterations of Gibbs sampling.

Table 1: Statistics for models used as examples.

Name	Docs	Tokens	Vocab	Topics
News	1.8M	76M	121k	1000
Blogs	13k	2.2M	90k	100
Parliament	540k	55M	52k	300

New York Times articles. Figure 3 shows three groupings of words for the middle-left topic in Figure 1: by document, by month of publication (e.g. May of 2005), and by desk (e.g. Editorial, Foreign, Financial). Instantaneous mutual information values are significantly smaller for the larger groupings, but the actual values are still larger than expected under the model. We are interested in measuring the degree to which word usage varies within topics as a function of both time and the perspective of the article. For example, we may expect that word choice may differ between opinion articles, which overtly reflect an author's views, and news articles, which take a more objective, factual approach.

We summarize each grouping by plotting the distribution of deviance scores for all topics. Results for all 1000 topics grouped by documents, months, and desks are shown in Figure 4.

[3] http://www.ldc.upenn.edu

[4] http://www.hansard-archive.parliament.uk/

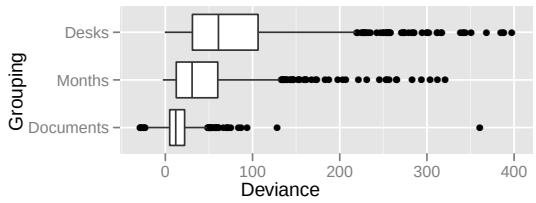

Figure 4: **News: Lack of fit correlates best with desks.** We calculate the number of standard deviations between the mean replicated discrepancy and the actual discrepancy for each topic under three groupings. Boxes represent typical ranges, points represent outliers.

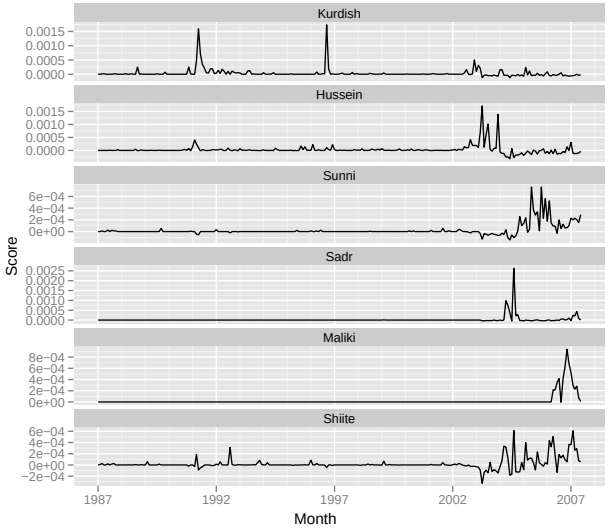

Figure 5: **News: Events change word distributions.** Words with the largest MI from a topic on Iraq's government are shown, with individual scores grouped by month.

Finally, we can analyze how individual words interact with groupings like time or desk. Figure 5 breaks down the per-word discrepancy shown in Figure 3 by month, for the words with the largest overall discrepancy. *Kurdish* is prominent during the Gulf War and the 1996 cruise missile strikes, but is less significant during the Iraq War. Individuals (*Hussein*, *Sadr*, and *Maliki*) move on and off the stage.

Political blogs. The CMU 2008 political blog corpus consists of six blogs, three of which supported Barack Obama and three of which supported John McCain. This corpus has previously been considered in the context of aspect-based topic models (Ahmed and Xing, 2010) that assign distinct word distributions to liberal and conservative bloggers. It is reasonable to expect that blogs with different political leanings will use measurably different language to describe the same themes, suggesting that there will be systematic deviations from a multinomial hypothesis of exchangeability of words within topics. Indeed, Ahmed and Xing obtained improved results with such a model. Figure 6 shows the distribution of standard deviations from the mean replicated value for a set of 150 topics grouped by document, blog, and preferred candidate. Deviance is greatest for blogs, followed by candidates and then documents.

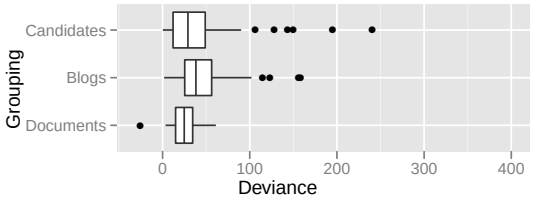

Figure 6: **Blogs: Lack of fit correlates more with blog than preferred candidate.** Grouping by preferred candidate has only slightly higher average deviance than by documents, but the variance is greater.

Grouping by blogs appears to show greater deviance from mean replicated values than grouping by candidates, indicating that there is further structure in word choice beyond a simple liberal/conservative split. Are these results, however, comparable? It may be that this difference is explained by the fact that there are six blogs and only

two candidates. To determine whether this particular assignment of documents to blogs is responsible for the difference in discrepancy functions or whether any such split would have greater deviance, we compared random groupings to the real groupings and recalculate the PPC. We generated 10 such groupings by permuting document blog labels and another 10 by permuting document candidate labels, each time holding the topics fixed. The average number of standard deviations across topics was 6.6 ± 14.4 for permuted "candidates" compared to 37.9 ± 39.2 for the real corpus, and 10.6 ± 12.9 for permuted "blogs" compared to 44.4 ± 29.6 for real blogs.

British parliament proceedings. The parliament corpus is divided into 305 volumes, each comprising about three weeks of debates, with between 600 and 4000 speeches per session. In addition to volumes, 10 Prime Ministers were in office during this period.

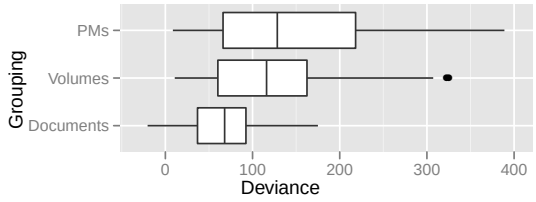

Figure 7: **Parliament: Lack-of-fit correlates with time (publication volume).** Correlation with prime ministers is not significantly better than with volume.

Grouping by prime minister shows greater average deviance than grouping by volumes, even though there are substantially fewer divisions. Although such results would need to be accompanied by permutation experiments as in the blog corpus, this methodology may be of interest to historians.

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. Figure 8 shows the most "mismatching" words for a topic with the most probable words *ships, vessels, admiralty, iron, ship, navy*, consistent with changes in naval technology during the Victorian era (that is, wooden ships to "iron clads"). Words that occur more prominently in the topic (*ships, vessels*) are also variable, but more consistent across time.

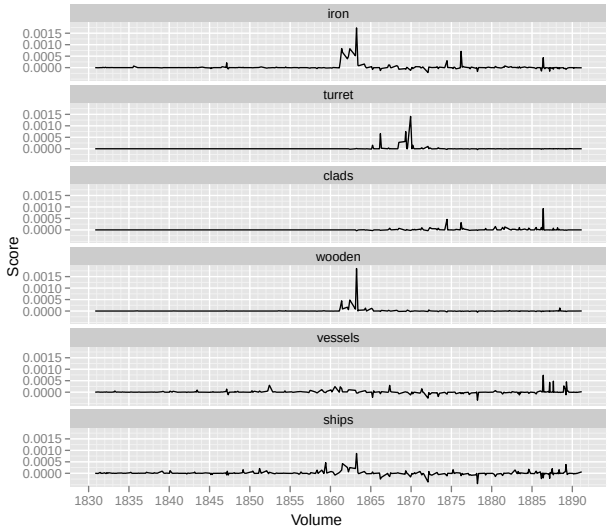

Figure 8: **Parliament: *iron-clads* introduced in 1860s.** High probability words (*ships, vessels*) are variable, but show less concentrated discrepancy than *iron, wooden*.

5 Calibration on Synthetic Data

A posterior predictive check asks "do observations sampled from the learned model look like the original data?" In the previous sections, we have considered PPCs that explore variability within a topic on a per-word basis, measure discrepancy at the topic level, and compare deviance over all topics between groupings of documents. Those results show that the PPC detects deviation from multinomial assumptions when it exists: as expected, variability in word choice aligns with known divisions in corpora, for example by time and author perspective. We now consider the opposite direction. When documents are generated from a multinomial topic model, PPCs should not detect systematic deviation.

We must also distinguish between lack of fit due to model misspecification and lack of fit due to approximate inference. In this section, we present synthetic data experiments where the learned model is precisely the model used to generate documents. We show that there is significant lack of fit introduced by approximate inference, which can be corrected by considering only parts of the model that are well-estimated.

We generated 10 synthetic corpora, each consisting of 100,000 100-word documents, drawn from 20

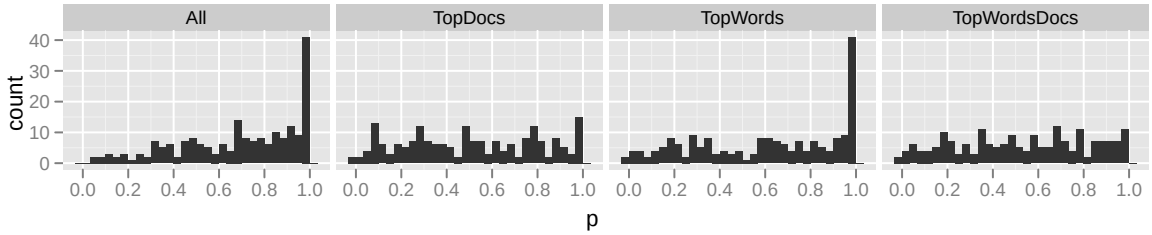

Figure 9: **Replicating only documents with large allocation in the topic leads to more uniform p-values.** p-values for 200 topics estimated from synthetic data generated from an LDA model are either uniform or skewed towards 1.0. Overly conservative p-values would be clustered around 0.5.

topics over a vocabulary of 100 terms. Hyperparameters for both the document-topic and topic-term Dirichlet priors were 0.1 for each dimension. We then trained a topic model with the same hyperparameters and number of topics on each corpus, saving a Gibbs sampling state.

We can measure the fit of a PPC by examining the distribution of empirical p-values, that is, the proportion of replications $\boldsymbol{w}^{rep}$ that result in discrepancies less than the observed value. p-values should be uniformly distributed on $(0, 1)$. Non-uniform p-values indicate a lack of *calibration*. Unlike real collections, in synthetic corpora the range of discrepancies from these replicated collections often includes the real values, so p-values are meaningful. A histogram of p-values for 200 synthetic topics after 100 replications is shown in the left panel of Figure 9.

PPCs have been criticized for reusing training data for model checking. For some models, the posterior distribution is too close to the data, so all replicated values are close to the real value, leading to p-values clustered around 0.5 (Draper and Krnjajic, 2006; Bayarri and Castellanos, 2007). We test divergence from a uniform distribution with a Kolmogorov-Smirnov test. Our results indicate that LDA is not overfitting, but that the distribution is not uniform (KS $p < 0.00001$).

The PPC framework allows us to choose discrepancy functions that reflect the relative importance of subsets of words and documents. The second panel in Figure 9 sums only over the 20 documents with the largest probability of the topic, the third sums over all documents but only over the top 10 most probable words, and the fourth sums over only the top words and documents. This test indicates that the distribution of p-values for the subset *TopWords* is not uniform (KS $p < 0.00001$), but that a uniform distribution is a good fit for *TopDocs* (KS $p = 0.358$) and *TopWordsDocs* (KS $p = 0.069$).

6 Conclusions

We have developed a Bayesian model checking method for probabilistic topic models. Conditioned on their topic assignment, the words of the documents are independently and identically distributed by a multinomial distribution. We developed a realized discrepancy function—the mutual information between words and document indices, conditioned on a topic—that checks this assumption. We embedded this function in a posterior predictive check.

We demonstrated that we can use this posterior predictive check to identify particular topics that fit the data, and particular topics that misfit the data in different ways. Moreover, our method provides a new way to visualize topic models.

We adapted the method to corpora with external variables. In this setting, the PPC provides a way to guide the modeler in searching through more complicated models that involve more variables.

Finally, on simulated data, we demonstrated that PPCs with the mutual information discrepancy function can identify model fit and model misfit.

Acknowledgments

David M. Blei is supported by ONR 175-6343, NSF CAREER 0745520, AFOSR 09NL202, the Alfred P. Sloan foundation, and a grant from Google. David Mimno is supported by a Digital Humanities Research grant from Google. Arthur Spirling and Andy

Eggers suggested the use of the Hansards corpus.

References

Amr Ahmed and Eric Xing. 2010. Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In *EMNLP*.

Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Asynchronous distributed learning of topic models. In *NIPS*.

M.J. Bayarri and M.E. Castellanos. 2007. Bayesian checking of the second levels of hierarchical models. *Statistical Science*, 22(3):322–343.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *ICML*.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January.

Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22*, pages 288–296.

Gabriel Doyle and Charles Elkan. 2009. Accounting for burstiness in topic models. In *ICML*.

David Draper and Milovan Krnjajic. 2006. Bayesian model specification. Technical report, University of California, Santa Cruz.

Jacob Eisenstein and Eric Xing. 2010. The CMU 2008 political blog corpus. Technical report, Carnegie Mellon University.

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

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235.

Matthew Hoffman, David Blei, and Francis Bach. 2010. Online learning for latent dirichlet allocation. In *NIPS*.

Wei-Hao Lin, Eric Xing, and Alexander Hauptmann. 2008. A joint topic and perspective model for ideological discourse. In *PKDD*.

Qiaozhu Mei and ChengXiang Zhai. 2006. A mixture model for contextual text mining. In *KDD*.

David Mimno and Andrew McCallum. 2007. Organizing the OCA: learning faceted subjects from a library of digital books. In *JCDL*.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Michael J. Paul, ChengXiang Zhai, and Roxana Girju. 2010. Summarizing contrastive viewpoints in opinionated text. In *EMNLP*.

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

D. Rubin. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172.

Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *ICML*.