2011

Correlations and Anticorrelations in LDA Inference

Alexandre Passos
Hanna M. Wallach, *University of Massachusetts - Amherst*
Andrew McCallum

Correlations and Anticorrelations in LDA Inference

Alexandre Passos
Department of Computer Science
University of Massachusetts Amherst
apassos@cs.umass.edu

Hanna M. Wallach
Department of Computer Science
University of Massachusetts Amherst
wallach@cs.umass.edu

Andrew McCallum
Department of Computer Science
University of Massachusetts Amherst
mccallum@cs.umass.edu

1 Introduction

Inference of the document-specific topic distributions in latent Dirichlet allocation (LDA) [2] and decoding in compressed sensing [3] exhibit many similarities. Given a matrix and a noisy observed vector, the goal of both tasks is to recover a sparse vector that, when combined with the matrix, provides a good explanation of the noisy observed data. In the case of LDA, the matrix corresponds to the topic-specific distributions over words, the noisy observed vector corresponds to the observed word frequencies for a single document, and the sparse vector corresponds to the document-specific distribution over topics for that document. In the scenario typically considered in the compressed sensing literature (i.e., a small, dense observed vector and a large, very sparse latent vector) the latent structure can be recovered exactly provided the matrix in question satisfies the restricted isometry property [4]. Satisfying this property means that no row in the matrix can be reconstructed from a sparse, linear combination of the other rows. Even though it is infeasible to test whether an arbitrary matrix satisfies the restricted isometry property, it is possible to use intuitions about this property, along with theorems about random matrices, to design improved compressed sensing systems and to prove theorems about the situations in which compressed sensing will work and why.

In this paper, we present preliminary work on identifying an analogue of the restricted isometry property for LDA, along with its effect on inference of the document-specific topic distributions. This work is based on the following observation: If a document contains occurrences of some word type that can be well-explained by multiple topics (e.g., the word "neural" in a corpus of NIPS papers, which could be explained by either topics on neural networks or topics on neuroscience), the model structure and associated inference procedures will almost always force the inferred topic distribution for that document to exhibit a strong preference for only one of these topics. Not only is this behavior required by the structure of optimal solutions to the LDA inference problem (see lemma 4 of Sontag and Roy [6]) but, intuitively, it is also this behavior that permits inference of the topic-specific word distributions: By explaining all document-specific occurrences of a word type with only one topic and searching for sparse document–topic distributions, the latent topics can "move apart" and eventually assign high probabilities to words that exhibit within-topic, but not across-topic, co-occurrences.

We illustrate our claims using Blei and Lafferty's correlated topic model [1]. This model is well-known to assign higher probabilities to held-out (i.e., previously unseen) documents than LDA, while simultaneously producing lower-quality topics, as judged by human evaluators [5].

2 Polysemous words and spurious correlations

The primary intuition underlying this paper is that topics whose high probability words overlap with those of other topics cause problems when inferring document-specific topic distributions. Such overlapping topics are very unlikely a priori: Given a large vocabulary, a small number of topics, and a sparse Dirichlet prior over the topic–word distributions, it is rare that multiple topic–word distributions sampled from this prior will assign a high probability to any given word type. Since the topic–word distributions are sampled independently, the probability that a given word type is among the highest probability words for any two topics decreases with the vocabulary size. Despite this, overlapping topics are common in learned LDA models (and have been used as evidence of polysemy [7]). Furthermore, this is problematic because most inference algorithms for LDA, and even the structure of optimal solutions, are systematically biased in these circumstances.

Given that polysemous word types exist, and that overlapping topics are common in learned LDA models, it is important to explore the resultant implications for inference of the document-specific topic distributions. Lemma 4 of Sontag and Roy [6] states that, for any document, an optimal solution to the problem of learning the optimal document–topic distribution will result in most word tokens being explained by a single topic. Even if some subset of the tokens in that document could be explained by multiple topics, the inference procedure will use as few of these topics as possible. Although this behavior could be prevented by using samples from the posterior distribution, rather than finding the optimal solution, MCMC methods for LDA are well-known to mix poorly and therefore most samples will be drawn from the region around a mode (i.e., an optimal solution).

An important effect of topics with overlapping high probability words is that, all else being equal, even if the topics are perfectly uncorrelated in the prior distribution, they will be anti-correlated in the posterior document–topic distributions learned by any reasonable inference algorithm for LDA. In this paper, we refer to such anti-correlations between overlapping topics as "spurious correlations".

Although it is less likely, the converse can also occur: pairs of topics with no overlapping high probability words will occur together in learned LDA models more often than topics with significant overlap, and will appear to be correlated in the inferred document-specific topic distributions.

Finally, it is worth noting that the relative absence of word types that occur with a high probability in multiple topics is a stronger requirement than the restricted isometry property. This suggests that topic models with assumptions closer to those of compressed sensing might be worth investigating.

2.1 Implications for correlated LDA

Since overlapping topics in LDA will induce spurious correlations in the learned document–topic distributions, it is natural to investigate this effect in a topic model which explicitly tries to learn correlations between topics, such as the correlated topic model (also known as correlated LDA) [1].

In LDA, the topic-specific distributions over words are drawn independently from a Dirichlet prior, as are the document-specific distributions over topics. Each word token is assumed to have been generated by first drawing a topic from the document-specific distribution over topics and then drawing a word from the corresponding topic-specific distribution over words. Since real-world topics are generally not independent and some pairs of topics are more likely to co-occur than others a priori, Blei and Lafferty [1] proposed correlated LDA. In this model, each document's distribution over topics is drawn from a logistic normal distribution, where the covariance matrix that parameterizes this logistic normal distribution (along with its mean) captures any correlations between topics.

Our main hypotheses are as follows:

- Learning a correlated topic model using data generated from LDA will find spurious anti-correlations when some topics' high probability words overlap with those of other topics.
- In a correlated topic model learned from real-world data, some of the largest observed anti-correlations will be between topics with overlapping high probability words, while some of the largest anti-correlations will be between topics without significant overlap.

If these hypotheses are true, then they explain the fact that correlated LDA results in higher held-out probabilities than LDA even when the inferred topics are less-interpretable. Since spurious correlations depend only on the structure of the topics (i.e., whether they have overlapping high probability

words), the inferred document–topic distribution for any held-out document containing words that can be well-explained by multiple topics will exhibit a strong preference for one of these topics, again inducing spurious correlations. The logistic normal distribution assigns a higher probability to correlated document–topic distributions than does the Dirichlet distribution. Therefore, correlated LDA should assign a higher probability to held-out data simply because it can predict its own bias.

These observations have broad implications for other topic models: in practice, none of the proposed priors for topic–word distributions can capture the that fact that polysemous words should occur with a high probability in multiple topics a priori. Additionally, the model structures and resultant inference algorithms exhibit biases in the presence of such topics. This problem is a result of the admixture property of LDA. In summary, the work presented in this paper suggests that interactions between the inferred document–topic and topic–word distributions are not well understood.

3 Experiments with correlated LDA

To test our hypotheses, we perform experiments on both synthetic and real-world data.

3.1 Synthetic data

The first hypothesis is that learning a correlated topic model using data generated from LDA will find spurious anti-correlations when topics have overlapping high probability words.

As an attempt to falsify this hypothesis, we performed the following experiment:

1. We generated four topics over forty word types, according to the following configurations:
   (a) **Treatment:** two topics that each assign high probability (in a 100-to-1 ratio) to ten unique words and low probability to the rest of the vocabulary, and two topics that each assign high probability to five unique words each and five shared words.
   (b) **Control:** four topics that each assign high probability to ten unique words.
2. We then generated 1000 synthetic documents (20 tokens each) using document–topic distributions drawn from a symmetric Dirichlet with a concentration parameter equal to two.
3. Finally, we fit a correlated topic model on each of these data sets using Blei's `ctm-c` implementation [1] (with default hyperparameters) and examined the covariance matrices.[1]

Interpreting these covariance matrices is not trivial because each one parameterizes a distribution over log probabilities (rather than probabilities) and they are 3×3 rather than 4×4; however, there are some clear differences. The covariance matrix for the **treatment** case (overlapping topics) is:

$$\begin{bmatrix} 1.5 & -0.2 & -0.6 \\ -0.2 & 1.5 & -0.7 \\ -0.6 & -0.7 & 1.0 \end{bmatrix}$$

The covariance matrix for the **control** case (non-overlapping topics) is:

$$\begin{bmatrix} 3.0 & 0.6 & 0.7 \\ 0.6 & 2.6 & 0.7 \\ 0.7 & 0.7 & 2.8 \end{bmatrix}$$

As expected, the covariance matrix for the **treatment** case contains spurious anti-correlations.

A possible criticism of these experiments is that the overlapping topics can be seen as combinations of some other set of "underlying" correlated topics. For example, one underlying topic could account for the shared words, with other underlying topics accounting for the other high probability words. However, the underlying topics would then be *positively* correlated, rather than anti-correlated.

3.2 Experiments on the NIPS corpus

The second hypothesis is that a correlated topic model, when fit on real-world data, will exhibit relatively large observed anti-correlations between topics with overlapping high probability words. Meanwhile, the largest correlations will be between topics that do not overlap significantly.

[1] Note that the true topics were not used when fitting the correlated topic model.

We fit a correlated topic model on the NIPS corpus.[2] We expect any topic model fit on NIPS papers to contain at least one topic consisting of words related to neural networks and at least one topic consisting of words neuroscience-related words. Furthermore, both sets of topics should assign relatively high probabilities to some common words, such as "neural," "connection," "activity," etc. Therefore, if our hypotheses are true, some of the largest anti-correlations observed in a correlated topic model fit on NIPS papers should be between neural network topics and neuroscience topics.

We fit a correlated topic model with twenty topics (again using `ctm-c` with default settings). We filtered out a small list of stopwords and downcased all words, but otherwise performed no stemming or vocabulary reduction. The inferred topics are in table 1. To extract meaningful correlations between topics we computed the empirical covariance matrix between the probabilities (not the unnormalized log probabilities) of the topics for each document. Although this does not provide accurate estimates of correlations between topics, the extremes of this distribution—that is, the most- and least-correlated topics—should be reasonably stable. The strongest anti-correlations found were, in order, between topics 8 and 15 and topics 6 and 8. Topics 6 and 8 are, respectively, a neuroscience and a neural network topic. The strongest correlations were found between topics 2 and 13, topics 5 and 6, and topics 7 and 9. None of these pairs overlap except for a few single-letter words.

4 Conclusions and future work

Our analyses suggest that the structural assumptions underlying LDA-based topic models are not well-understood. Understanding these assumptions, and the limitations they induce, can help practitioners better use topic models and help researchers design models with alternative assumptions.

One possible avenue for future work is investigating the impact that the analyses presented in this paper have on inference of the topic-specific distributions over words. Finally, requiring that topics have non-overlapping high probability words is a very strong requirement—stronger than the restricted isometry property. It is therefore necessary to investigate whether this is indeed the main structural property separating easy and hard inference problems in LDA-based topic models.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by IARPA via DoI/NBC contract #D11PC20152, and in part by NSF grant #SBE-0965436. The U.S. Government is authorized to reproduce and distribute reprint for Governmental purposes notwithstanding any copyright annotation thereon. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[1] D.M. Blei and J.D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

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

[3] E.J. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 1433–1452, 2006.

[4] E.J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[5] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, 2009.

[6] David Sontag and Daniel Roy. Complexity of inference in latent Dirichlet allocation. In *Neural Information Processing Systems*, 2011.

[7] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440, 2007.

[2] `http://www.cs.toronto.edu/~roweis/data.html`

Topic	High-probability words
000	speech we with are p state as this on model word recognition models s hmm system an context training from sequence using time l can
001	feature are as ith classification classifier this input features map pattern which an each patterns on class neural network j vector can information n k
002	n we r l f this v function as with o j p d m s where are on an c which it g
003	neuron with are this figure circuit analog as neurons on chip an neural input time output current voltage from synaptic at network model which system
004	network are this on as s it with rules neural learning an from networks which training rule set or one can not each input have
005	motion spike this direction at as cells we with cell are firing information s from on model time neurons j rate r an which stimulus
006	model are with visual cells from this as on cortex j which activity neurons at input orientation s cortical response cell between figure stimulus m
007	p we with data o y c w are as on this z k e d distribution n f model set training which function from
008	network training with on layer hidden this are networks input as neural error output was units set performance we net from weights were each number
009	data model are with we models this on as mixture p from m j variables which can each it probability set algorithm e em used
010	algorithm with on re k this problem as set data vector c algorithms function distance can kernel which training optimization j method vectors support an
011	image are imag as this we from with on object an each can s which figure it two at or d visual one using not
012	network state time as with system this are neural dynamics model we can an networks which it j b recurrent at s from on memory
013	we n with on e p f w m s k are c can this l if d as function h learning networks functions
014	learning with netw time are networks as on neural control algorithm error weight propagation this gradient training e problem an weights back figure was can
015	learning state s this with q on as time an are at it value policy r can action function reinforcement from which algorithm each
016	face on are with detection as we recognition network set training images from this an neural image ensemble each was faces view human used were
017	noise s we re with as e this time from data m on p model signal j mean an analysis r can using q gaussian
018	w we this matrix are linear k as j y with n an s on can gradient algorithm function which h z filter r from
019	input we this are units can an each space network as from unit which its figure output on hidden learning one has e our j

Table 1: A subset of the topics learned from the NIPS corpus.