The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Reliability of Perplexity to Find Number of Latent Topics

Asana Neishabouri, Michel C. Desmarais
Polytechnique Montreal, Canada

Abstract

The problem of finding the correct number of latent topics in Latent Dirichlet Allocation is typically addressed by using a so-called *wrapper* approach and optimizing over the perplexity measure. This problem can be considered a dimensionality reduction task. We investigate how popular methods from different fields determine the right number of latent factors to retain. We address the reliability of these methods under different conditions and under different characteristics of datasets.

In particular, we show that although perplexity is the favorite statistical method to choose the number of latent topics, it does not systematically outperform other methods under different matrix sparsity levels. We show that SVD-based methods and a well-known methods in psychometrics sometimes yield the greatest performances. We also show that we can take advantage of antithetical results across methods to estimate the reliability of the estimated number of latent topics.

Introduction

Finding the number of hidden factors is a common problem for a number of statistical and machine learning techniques that are deployed in fields such as information retrieval, psychology, and recommender systems. Interestingly, each field of study has its own methods of choice to solve this problem. Few studies borrow methods from outside their fields to investigate the reliability and performance of these methods within a single field. In the field of Topic Modeling, this problem translates to the task of finding the number of topics.

We investigate if, and how methods outside of the typical topic modeling studies can tackle this task. Experiments are conducted with synthetic data where we know the ground truth (number of topics) and, as a generative model, LDA is well suited for that purpose.

The results of the experiments surprisingly show that, under certain conditions, the linear methods show better estimate the number of topics than perplexity. We also find that sparsity has a key effect and even more important than α and β to find the correct number of topics (K) by the mentioned methods.

The rest of this paper is organized as follows. We first review LDA in more details and review the best known and most successful methods to find the correct number of latent factors in different fields. Then, we report the details of our experiments. Next, we discuss the results before concluding.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) was originally introduced by Blei, Ng, and Jordan (2003) and is arguably the most widely used method for topic modeling. It is a generative, probabilistic model of documents. This technique uses words distribution to cluster texts and discover latent topics from it. It is based on the assumptions that each document consists of a mixture of hidden topics and that each topic consists of a set of words, both of which follow a Dirichlet prior.

LDA has the following hyperparameters that have to be determined before the training phase:

- Alpha (α), is the Dirichlet prior of document-topic density. Higher values of alpha implies documents composed of more topics and lower values implies fewer topics per document.
- Beta (β), is the Dirichlet prior of document-topic density. A higher beta indicates that topics are composed of a large vocabulary, and a lower value implies smaller vocabularies per topic.
- K is the number of topics.

Griffiths and Steyvers (2004) suggest a value of $50/T$ for α, where T is the number of topics, and 0.1 for β. And consequently many studies such as (Jameel and Lam 2013; Shang and Chan 2010) as well as the “topicmodels” package (Grün and Hornik 2011) in R use of these values.

For finding the number of topic parameters K, most studies (Blei, Ng, and Jordan 2003; Su and Liao 2013, for eg.) use perplexity along with a wrapper technique to find the number that will essentially yield the most likely probability of the data given the hyperparameters and the training parameters derived. A fair number of studies have shown its effectiveness (Blei, Ng, and Jordan 2003; Vu, Li, and Law 2019; Henderson and Eliassi-Rad 2009; Cha and Cho 2012; Hoffman, Bach, and Blei 2010; Zhang et al. 2007).

To skip defining the number of topics in advance, Hierarchical Topic Modeling (HTM) was proposed and studied by a number of researchers (Griffiths et al. 2004; Blei, Griffiths, and Jordan 2010; Mimno, Li, and McCallum 2007; Paisley et al. 2014). However, other studies (Kang, Ma, and Liu 2012; Mao et al. 2012; Wallach, Mimno, and McCallum 2009) have shown that hierarchical topic modeling does not yield better results and argue that some of the cons of HTM methods are that they are heuristic-based and computationally expensive.

Dimension reduction methods

Let us now turn to alternative methods to address the problem of finding the number of latent topics that are inspired from dimension reduction and factor analysis.

In the social science and psychometrics areas, factor analysis techniques help decide the number of latent factors to retain from a dataset (Ledesma and Valero-Mora 2007). Among the best known, we find Kaiser's eigenvalue-greater-than-one rule, Parallel Analysis (PA), Cattell's Scree test, Minimum Average Partial test which is known as Velicer's MAP. They are reviewed below.

In areas of machine learning applications such as recommender systems, we find Singular Value Decomposition (SVD) based approaches such as Bi-Cross-Validation (BCV), which is known as a wrapper method, and the more recent Randomize-SVD (RSVD) (Neishabouri and Desmarais 2019)

In this paper we evaluate the method PA, which according to the psychometrics literature is a top performer (Zwick and Velicer 1986), alongside with SVD-Based methods and perplexity in finding the correct number of topics under different conditions of datasets. This choice corresponds to our assessment of the most promising alternatives to perplexity (Neishabouri and Desmarais 2019). Moreover, we investigate the reliability of these methods and compare them.

Parallel Analysis

Horn's Parallel analysis (PA) (Horn 1965) is a well-known technique in psychometrics that is almost ignored by the machine learning community. It relies on the correlation matrix of the observed variables of the original datasets, and multiple random datasets generated having the same size as the original dataset. A factor is retained if its associated eigenvalues of the correlation matrix of the original dataset are bigger than the mean or 95 percentile eigenvalues that are derived from the correlation matrices of the random generated datasets. The remaining factors are considered random noise.

In this article, mPA and cPA refer to the two variations of PA method that correspond respectively to the average and 95 percentile of the eigenvalues of the random datasets respectively.

Bi-Cross-Validation (BCV) of the SVD

Gabriel cross-validation (BCV.G) and Wold style cross-validation (BCV.W) are wrapper methods that rely on cross validating the singular value decomposition (SVD) to find the best rank of a matrix to truncate the SVD (Owen, Perry, and others 2009).

The singular value decomposition of a matrix is a well-known matrix factorization technique. It decomposes the original matrix, $\mathbf{A}$, into three matrices as below.

$$\mathbf{A} = \mathbf{U}_{m\times m}\,\mathbf{\Sigma}_{m\times n}\,\mathbf{V}^T_{n\times n}$$

where U and V are two eigenvector matrices that are orthonormal and called the left-singular vectors and right-singular vectors of $\mathbf{A}$ respectively and the matrix $\mathbf{\Sigma}$ is diagonal with non-negative real values.

Both the Gabriel-style (BCV.G) and Wold-style (BCV.W) cross validation variations consists in dividing the data into a training and a test set. Prediction is done with a truncated (lower rank) product of the factorization. The prediction error is measured as the sum of squares of residuals between the truncated SVD and the original matrix. Determining the number of LD relies on a comparison of the residual error over a random set of values for the test set.

In the Wold-style cross validation, the test set is a random set of values in the matrix. The difference between Wold- and Gabriel-style is that, for the later, the test set is "blocked". It holds out a certain number of rows and columns of a matrix simultaneously as a test set. BCV.G divides the rows of the matrix into k segments and the columns into h segments. The total number of folds are $k \times h$ which refer to the number of blocks. In each step, one of the blocks is considered as the test set and the remaining blocks are as the training set to reconstruct the original matrix. More details about this method are given in (Owen, Wang, and others 2016; Kanagal and Sindhwani 2010)

With large data sets free of missing values, Owen, Wang, and others (2016) report that BCV has a better result than other methods in the state of the art. Several studies indicate that the Wold-style cross validation provides a better result but is slower than Gabriel-style (Kanagal and Sindhwani 2010; Owen, Perry, and others 2009).

Randomized Singular Value Decomposition: RSVD

Randomized Singular Value Decomposition (RSVD) is briefly introduced in (Beheshti, Desmarais, and Naceur 2012) and developed further in (Neishabouri and Desmarais 2019). This method is similar to PA. It compares the singular values of the original matrix with the randomized matrix. The randomized matrix is a sample of the original matrix by selecting columns randomly with the same size.

The number of latent values is determined as the point of intersection of the two curves of the randomized and original matrices SVD singular values.

Perplexity

Perplexity is arguably the most popular metric in language modeling. It is based on the probability of the unseen test set, normalized by the number of words to evaluate the goodness of LDA model. The perplexity is defined as:

$$PP(\mathbf{W}) = P(\mathbf{W}_1\mathbf{W}_2, ..., \mathbf{W}_{\mathrm{m}})^{-1/m})$$

where PP and $\mathbf{W}$ refer to perplexity and word respectively, and $\mathbf{P}$ is the probability estimate assigned to document words. A model with a given number of topics that minimize perplexity value is considered an optimal model (Chen, Yao, and Yang 2016; Vu, Li, and Law 2019) .

Experiments and results

We conduct experiments to compare the methods outlined above over the task of inferring the number of topics used to generate synthetic documents generated with the LDA model. Specifics of each method for the experiments are:

- Parallel Analysis (PA): we use the "paran" library (Dinno 2018) in R. We examine the accuracy of the mean eigenvalue rule and the 95th percentile eigenvalue rule. We call them "mPA" and "cPA" respectively.
- Bi-Cross Validation: Wold-style cross validation with 5 folds and also Gabriel-style cross validation with 4 sub matrices (Perry 2015) which are the default of the library. We refer to them as BCV.W and BCV.G respectively.
- Randomized SVD (RSVD): implemented in R using algorithm 1 in (Neishabouri and Desmarais 2019).
- Perplexity: We use LDA with Gibbs sampling and perplexity functions in "topicmodels" library (Grün and Hornik 2011) in R.

PA, BCV and RSVD are linear methods to find the number of latent factors, while LDA is a non-linear method which uses perplexity to find the optimal number of latent factors to retain.

In our experiments we generate document-term matrices from a LDA model with known hyper-parameters. Using synthetic data is our choice of validation method because we know the ground truth of the generated data and we can control the different characteristics to investigate. We design some experiments that simulate the short texts such as reviews, comments, abstracts, tweets, etc. We aim to find the correct number of topics behind the generated datasets using the mentioned methods in order to compare their performance.

We explore the performance of methods over document-term datasets of size $= 250 \times 1000$, where the rows and columns refer to the number of documents and vocabulary respectively. We also consider two sets of priors (alpha and beta) to generate document-term matrices, namely set 1 $=$ $[\alpha = 0.6, \beta = 0.1]$ and set 2 $= [\alpha = 0.8, \beta = 0.6]$.

We study the methods' performance for different dataset characteristics such as the number of latent topics, level of sparsity (number of terms per document) for each set of priors. Hence, we generate 15 different datasets for each set of priors, and we further generate datasets for 3 sizes of latent topics, $K = [5, 10, 15]$ and that for different levels of sparsity using the number of terms per document such as $[50, 100, 200, 300, 400]$.

Datasets generation

To generate synthetic document-term matrices, we rely on the generative nature of LDA (Blei, Ng, and Jordan 2003).

Algorithm 1 shows the procedure and steps to perform our experiment per each dataset.

Results of the experiments over the different data sets are reported below.

Algorithm 1 Experiment Procedure

```
• Define size of documents and vocabulary
• Define $\alpha$, $\beta$ and $K$
• DTM ← Generate synthetic Document-Term dataset using LDA generative Process
• Estimate K using each of the linear method (DTM)
• Estimate K using Perplexity through following steps:
  1. Split DTM into five folds
  2. For each unique fold
  (a) Take the fold as a test
  (b) Take rest of the folds as a training set
  (c) For $K = 2 : 25$
      i. Fit LDA model using Gibbs sampling on the training set
      ii. Evaluate the fitted model on the test set using perplexity
      iii. Retain the evaluation score per each $K$
  (d) Take the average of the perplexity score for each $K$
  3. Return K with minimum average perplexity score
```

Experiment 1, $\alpha = 0.6$ and $\beta = 0.1$

In this experiment, we explore each method on the generated datasets with $\alpha = 0.6$ and $\beta = 0.1$. Figure 1 (top plot) displays the estimation of each method at different levels of sparsity (number of terms per documents) $= [50, 100, 200, 300, 400]$ and latent topics, $= [5, 10, 15]$.

For the further analysis, we compute loss of the estimated number of latent topics using the Mean Absolute Error (MAE):

$$\text{loss}_{\text{MAE}} = 1/n \sum_{i}^{n} |K_i - \hat{K}_i|$$

Where K_i is the real number of latent topics of data set i and $\hat{K}_i$ is the estimated number.

We discuss the results in the next section.

Experiment 2, $\alpha = 0.8$ and $\beta = 0.6$

This experiment illustrates the effect of higher priors in each of the mentioned method's estimation. We generate all the datasets with $\alpha = 0.8$ and $\beta = 0.6$. Figure 1 (bottom plot) shows the estimation of each method at different levels of sparsity and latent topics. The details are discussed in the following section.

Discussion and analysis of errors

According to Figure 1, we can conclude that dataset characteristics such as sparsity, number of topics, α and β play a crucial role for the capacity of the methods to find the number of latent topics.

As we could expect, the error estimation of K grows as the real value of K grows, and as the size of the documents drop from 400 to 50 terms. Unexpected is that the direction of errors is opposite when comparing perplexity with all other methods. These error trends are analyzed below.

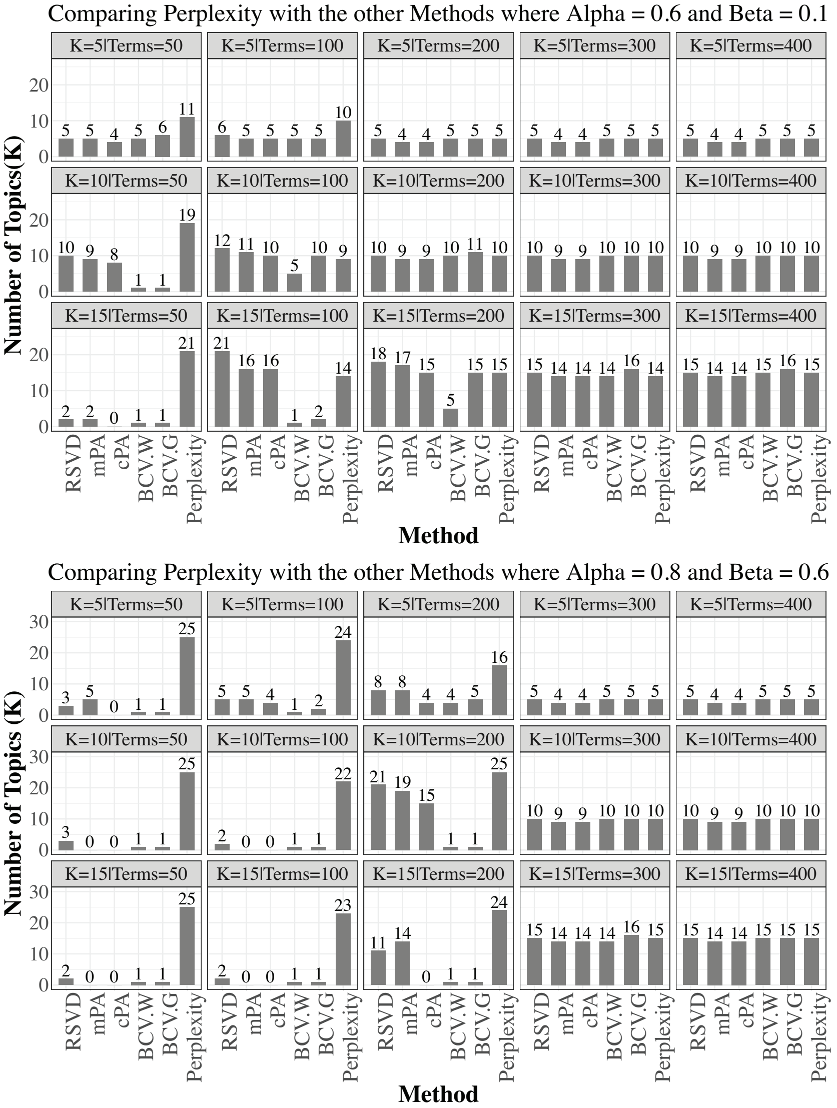

Figure 1: Estimation of the methods at each level of sparsity in the first (top) and second (bottom) experiments. In each panel, K and Terms in the figure refer to the number of topics and terms per document respectively.

Table 3 shows the ratio of correct and over/under estimation for each method averaged over all data sets. We find that the linear models RSVD and BCV often show a higher capability to estimate the right number of topics than perplexity. Moreover, we see that although BCV.W has the same accuracy as perplexity, it has a bias in the opposite direction. Perplexity overestimates the number of latent topics whereas BCV.W underestimate it. More on this trend below. Note that for the two plots in Figure 1, a bias correction is applied to mPA and cPA

Figure 1 also illustrates that all the methods have a higher loss with higher sparsity and higher priors (α, β).

In order to examine the effect of these parameters, we compute average loss of all the methods under each condition. Figure 2 displays the relation between average loss of all the methods and datasets characteristics such as sparsity (Terms per Documents), and each set of priors. We find that by increasing sparsity, all methods have a higher loss on average. Moreover, it shows that there is a higher average loss where priors have a higher value. We also can see that where there is a less sparsity, hyperparameters α and β do not affect the results significantly. Which means that if we could control the sparsity we can ignore the effect of Alpha and Beta values.

Another interesting point of the two plots in Figure 1 is that, with increased sparsity, the perplexity method overestimate whereas the other methods underestimate the number of topics. In order to investigate the behavior of perplexity with respect to the other methods, we define another variable as “status” that tells us whether each method over/underestimate if the estimations differ more than one from the correct latent topics. Table 1 shows the number of

Table 1: Perplexity Overestimation table

	Overestimate	Not_Overestimate
Opposite_Direction	9.00	0.00
Same_Direction	4.00	17.00

Table 2: Perplexity Overestimation Odds Ratio and Confidence Interval

Odds Ratio	Lower Limit CI	Upper Limit CI
73.89	3.581	1524.226

Table 3: Accuracy and over/under estimation of each method

Methods	Correct	Overestimate	Underestimate
RSVD	0.57	0.20	0.23
mPA	0.53	0.30	0.17
cPA	0.57	0.17	0.27
BCV.W	0.47	0.00	0.53
BCV.G	0.47	0.17	0.37
Perplexity	0.47	0.43	0.10

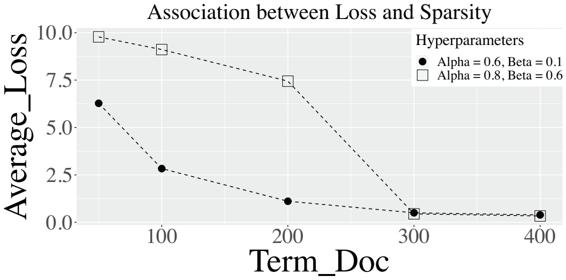

Figure 2: Association between sparsity, hyperparameters and loss

times perplexity overestimate in the same or opposite direction of the other methods.

To compare the behavior of perplexity with the other methods and assess if the overestimation of perplexity with respect to the other methods estimations is significant or not, we compute the odds ratio alongside the 95% confidence interval (CI) of perplexity. In order to avoid singular odds ratios, we make the Laplace correction, a kind of prior, and add 0.5 to all the values in table 1. Table 2 shows that the odds ratio of perplexity overestimation in opposite direction of the other method is 73.89 times greater than when is not overestimating and the CI also indicate that the odds ratio is statistically significant since it does not include 1.

It also worth mentioning that perplexity can overestimate even more where $\alpha = 0.8$ and $\beta = 0.6$, but since to avoid time consumption we set a range of $K = 2 : 25$ to evaluate the LDA model using perplexity, it shows the maximum one.

Conclusion and future work

We tackled the problem of finding the number of topics with well-known linear methods from the other fields that have not been utilized before. We showed that despite the fact these methods are within a linear framework and under certain conditions, some of them have a better performance than the commonly used perplexity measure to find the number of topics. We also show the boundaries where perplexity, as well as the other methods, become subject to unreliable estimations. We show that the performance deteriorates sharply as dataset sparsity and priors increase.

The boundaries have the interesting characteristic that the perplexity measure overestimates the number of latent dimensions, whereas the other methods understimates them. This leads to an interesting indicator that all methods are providing unreliable estimates.

The experiment results also corroborates the finding that LDA performs poorly with short texts (Li et al. 2019), and in particular that the commonly used perplexity measure to derive the number of topics overestimates this parameter with high sparsity.

An important limitation of this work is that we do not have a analytical explanation for the different behavior of the methods.

Current research in each domain mostly focuses on specific technical approaches and evaluation that are known within the field which makes it difficult to conclude that if the achieved results are actually the best that could be and reliable. Our experiments show the necessity of engaging to more multidisciplinary methods in order to be aware of the reliability of an evaluation approach despite the popularity. Our experiments and contribution could lead to a more reliable and accurate estimation of the number of topics.

References

Beheshti, B.; Desmarais, M. C.; and Naceur, R. 2012. Methods to find the number of latent skills. *International Educational Data Mining Society*.

Blei, D. M.; Griffiths, T. L.; and Jordan, M. I. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)* 57(2):7.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Cha, Y., and Cho, J. 2012. Social-network analysis using topic models. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 565–574. ACM.

Chen, Q.; Yao, L.; and Yang, J. 2016. Short text classification based on lda topic model. In *2016 International Conference on Audio, Language and Image Processing (ICALIP)*, 749–753. IEEE.

Dinno, A. 2018. *paran: Horn's Test of Principal Components/Factors*. R package version 1.5.2.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* 101(suppl 1):5228–5235.

Griffiths, T. L.; Jordan, M. I.; Tenenbaum, J. B.; and Blei, D. M. 2004. Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems*, 17–24.

Grün, B., and Hornik, K. 2011. topicmodels: An R package for fitting topic models. *Journal of Statistical Software* 40(13):1–30.

Henderson, K., and Eliassi-Rad, T. 2009. Applying latent dirichlet allocation to group discovery in large graphs. In *Proceedings of the 2009 ACM symposium on Applied Computing*, 1456–1461. ACM.

Hoffman, M.; Bach, F. R.; and Blei, D. M. 2010. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, 856–864.

Horn, J. L. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika* 30(2):179–185.

Jameel, S., and Lam, W. 2013. An n-gram topic model for time-stamped documents. In *European Conference on Information Retrieval*, 292–304. Springer.

Kanagal, B., and Sindhwani, V. 2010. Rank selection in low-rank matrix approximations: A study of cross-validation for nmfs. In *Proc Conf Adv Neural Inf Process*, volume 1, 10–15.

Kang, J.-H.; Ma, J.; and Liu, Y. 2012. Transfer topic modeling with ease and scalability. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, 564–575. SIAM.

Ledesma, R. D., and Valero-Mora, P. 2007. Determining the number of factors to retain in efa: An easy-to-use computer program for carrying out parallel analysis. *Practical assessment, research & evaluation* 12(2):1–11.

Li, J.; Huang, G.; Fan, C.; Sun, Z.; and Zhu, H. 2019. Key word extraction for short text via word2vec, doc2vec, and textrank. *Turkish Journal of Electrical Engineering & Computer Sciences* 27(3):1794–1805.

Mao, X.-L.; Ming, Z.-Y.; Chua, T.-S.; Li, S.; Yan, H.; and Li, X. 2012. Sshlda: a semi-supervised hierarchical topic model. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 800–809. Association for Computational Linguistics.

Mimno, D.; Li, W.; and McCallum, A. 2007. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th international conference on Machine learning*, 633–640. ACM.

Neishabouri, A., and Desmarais, M. C. 2019. Investigating methods to estimate the number of latent dimensions under different assumptions and data characteristics. Technical report.

Owen, A. B.; Perry, P. O.; et al. 2009. Bi-cross-validation of the svd and the nonnegative matrix factorization. *The annals of applied statistics* 3(2):564–594.

Owen, A. B.; Wang, J.; et al. 2016. Bi-cross-validation for factor analysis. *Statistical Science* 31(1):119–139.

Paisley, J.; Wang, C.; Blei, D. M.; and Jordan, M. I. 2014. Nested hierarchical dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(2):256–270.

Perry, P. O. 2015. *bcv: Cross-Validation for the SVD (Bi-Cross-Validation)*. R package version 1.0.1.

Shang, L., and Chan, K.-P. 2010. A temporal latent topic model for facial expression recognition. In *Asian Conference on Computer Vision*, 51–63. Springer.

Su, J., and Liao, W.-P. 2013. Latent dirichlet allocation for text and image topic modeling.

Vu, H. Q.; Li, G.; and Law, R. 2019. Discovering implicit activity preferences in travel itineraries by topic modeling. *Tourism Management* 75:435–446.

Wallach, H. M.; Mimno, D. M.; and McCallum, A. 2009. Rethinking lda: Why priors matter. In *Advances in neural information processing systems*, 1973–1981.

Zhang, H.; Qiu, B.; Giles, C. L.; Foley, H. C.; and Yen, J. 2007. An lda-based community structure discovery approach for large-scale social networks. In *2007 IEEE Intelligence and Security Informatics*, 200–207. IEEE.

Zwick, W. R., and Velicer, W. F. 1986. Comparison of five rules for determining the number of components to retain. *Psychological bulletin* 99(3):432.