Search for K: Assessing Five Topic-Modeling Approaches to 120,000 Canadian Articles

Dr. Qiang Fu
Department of Sociology
The University of British Columbia
Vancouver, BC, Canada
Email: qiang.fu@ubc.ca

Yufan Zhuang
Data Science Institute
Columbia University
New York, NY, USA
Email: yufan.zhuang@columbia.edu

Jiaxin Gu
Department of Sociology
The University of British Columbia
Vancouver, BC, Canada
Email: gujiaxinsoci@gmail.com

Dr. Yushu Zhu
Urban Studies Program and School of Public Policy
Simon Fraser University
Vancouver, BC, Canada
Email: yushu_zhu@sfu.ca

Huihui Qin
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hong Kong, China
Email: huihui.qin@connect.polyu.hk

Dr. Xin Guo
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hong Kong, China
Email: x.guo@polyu.edu.hk

***Abstract*— Topic modeling has been an important field in natural language processing (NLP) and recently witnessed great methodological advances. Yet, the development of topic modeling is still, if not increasingly, challenged by two critical issues. First, despite intense efforts toward nonparametric/post-training methods, the search for the optimal number of topics K remains a fundamental question in topic modeling and warrants input from domain experts. Second, with the development of more sophisticated models, topic modeling is now ironically been treated as a black box and it becomes increasingly difficult to tell how research findings are informed by data, model specifications, or inference algorithms. To address these issues, we employ five training methods (Latent Semantic Analysis, Latent Dirichlet Allocation, Principal Component Analysis, Factor Analysis, Non-negative Matrix Factorization) to identify discussion topics based on about 120,000 newspaper articles retrieved from three major Canadian newspapers (Globe and Mail, Toronto Star, and National Post) since 1977. The optimal topics are then assessed using three measures: coherence statistics, held-out likelihood, and graph-based dimensionality selection. Findings from this research not only complement important advances in topic modeling and but provide insights into the choice of optimal discussion topics in social science research.**

Keywords—Topic Modeling, Natural Language Processing, Social Science, Optimal Number of Topics

I. INTRODUCTION

The past two decades have witnessed an explosion in methods, algorithms and tools designed to identify discussion topics in automated text analysis. Noteworthy among these research efforts, the Latent-Dirichlet-Allocation (LDA) approach assumes a Dirichlet prior distribution assigning a specific set of topics to each document, based on a fixed number (K) of topics. By incorporating both observed and latent variables, this Bayesian generative method allows for latent processes to capture similarities among sets of observations and thus results in a more precise assignment of topics to documents (and words to documents) [1]. While this method has been further developed to specify the number of optimal discussion topics based on a nonparametric Bayesian model [2], in practice the ultimate decision on the choice of K still relies on significant input from domain experts. In a more recent review of data analysis with latent models, Blei highlights a tension between orthodox Bayesian thinking and model criticism [3]. While the former attempts to integrate all possible sources of uncertainties in a more complex mixture or "super" models, the latter tries to tell whether the essence of the data has been captured by model specification and/or parameter inference. Yet, model criticism is becoming increasingly challenging with the adoption and proliferation of latent models in that we do not necessarily know whether the data, model specification, or inferential algorithms plays a more significant part in shaping the (approximate) posterior. In response to these issues, this research uses various approaches to assess the choice of K via different training methods, where model specification and inferential algorithms play different roles in shaping research findings.

II. FIVE APPROACHES TO TOPIC MODELING

A. Term Frequency–Inverse Document Frequencies

To apply topic-modeling methods, we represent a large corpus of text using a document-word matrix X, where each column corresponds to a document and each row corresponds to a word [4]. Since a word's frequency in a corresponding document cannot suggest the word's relative importance in the whole corpus, elements of the document-word matrix are often weighted by term frequency–inverse document frequencies (tf-idf) [5]. One way to calculate the tf-idf weight $w_{t,d}$ associated with a term t and a document d is as follows [6],

$$w_{t,d} = tf_{t,d} \times \log \frac{N}{df_t}$$

where $tf_{t,d}$ is a term t's frequency in a document, N is the total number of documents, and df_t is the total number of documents containing the term t. Clearly, $w_{t,d}$ increases if a term has a higher frequency in a document but such increase is offset by the term's prevalence across all documents in a text corpus. This tf-idf weight thus tends to filter out common words or stop-words which appear to be popular in virtually all documents.

B. Latent Semantic Analysis

To guide our assessment of different approaches to topic modeling, we next discuss methodological details of the five models being adopted in this research. Based on singular value decomposition of the document-word matrix, latent semantic analysis (LSA) has long been adopted by scholars from different disciplines to identify topics and themes contained in text corpus [7]. This is achieved by providing a low-rank approximation to the previously defined word-document matrix X [8]. To understand how LSA works, we have its singular value decomposition (SVD) of X as:

$$X = U\Sigma V^T,$$

where both U and V are orthogonal matrices and Σ is a diagonal matrix. To further explore these three matrices, we first note that the square matrix XX^T contains all dot products denoting the correlation between any two word vectors across all documents, and X^TX contains all dot products denoting the correlation between any two document vectors. And we have:

$$U^T XX^T U = \Sigma\Sigma^T \text{ and } V^T X^T XV = \Sigma^T\Sigma, \text{ or}$$

$$XX^T = U\Sigma\Sigma^T U^T \text{ and } X^T X = V\Sigma^T\Sigma V^T.$$

In other words, XX^T and X^TX have the same eigenvalues expressed by $\Sigma\Sigma^T$ (or, equally by $\Sigma^T\Sigma$), and their eigenvectors are contained in U and V, respectively. The number of singular values in Σ suggests the rank of X, or the number of topics in the current research setting, while the values of these singular values suggests the relative importance of these topics. For a space spanned by singular vectors corresponding to these singular values (i.e., topics), the coordinates of a word i across all topics are denoted by the i^{th} row of U and the coordinates of a document j across all topics are denoted by the j^{th} column of V^T. The corresponding loadings of all words on the k^{th} topic are given by elements in the kth columns of U; and the corresponding loadings of all documents on the k^{th} topic are given by elements in the kth rows of V^T. While topics identified by LSA can be viewed as clusters of words and/or documents once they are projected to a "semantic space", we use columns of U to denote topics (and their corresponding relations with words). If the values of singular values are small or below a certain threshold specified by researchers, it is possible to remove these singular values and achieve a low-rank approximation [9].

C. Principal Component Analysis

The idea of principal component analysis (PCA) is very similar to that of SVD [10]. For the document-word matrix X, PCA tries to project the data to orthogonal directions so that distinctive features from the data can be retained as much as possible. In other words, if the covariance matrix associated with X is given by XX^T, PCA is looking for a projection matrix P such that after the projection the covariance matrix Y^TY of the resulted new document-word matrix Y=PX has the largest variance in these projection directions. Yet, one constraint in the search for P is that these projection directions suggested by P should be basis vectors and orthogonal to each other. Otherwise, the direction associated with the second largest variance will be always parallel to or even overlap with that associated with the largest variance (and so forth for the remaining directions), which provides little information of the data. As a consequence, the off-diagonal elements (i.e., covariance) of Y^TY should be zero and PCA essentially deals with an issue of optimization with a constraint. We have:

$$Y^TY = (\mathrm{PX})(\mathrm{PX})^T = PXX^TP^T = D$$

where D should be a diagonal matrix. Related to our discussion on SVD, if we rank eigenvectors $z_1, z_2, \cdots z_n$ of XX^T and form a new matrix $Z = \begin{pmatrix} z_1 & z_2 & \cdots & z_n \end{pmatrix}$ and let:

$$Z^T XX^T Z = \Sigma^T\Sigma = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} \quad (1).$$

D will be a diagonal matrix if we make $P=Z^T$. Therefore, the matrix containing all the eigenvectors of XX^T provides the loadings of all words on any topic and a solution to the application of PCA to topic modeling. The optimization issue also corresponds to the maximization of $z_i^T XX^T z_i$ when $z_i^T z_i = 1$. If we take the derivative of $z_i^T XX^T z_i - \lambda z_i^T z_i$ with respect to z_i, we have $(XX^T - \lambda \mathrm{I})z_i = 0$ and z_i must be an eigenvector of XX^T. To summarize, the relation between LSA and PCA is similar to that between maximum likelihood estimation and ordinary least squares estimation in linear

regression settings: they appear to follow different principles yet (sometimes) yield the same result. Nevertheless, these two methods differ from each other in terms of computing: the calculation involving covariance matrices can be demanding when observations and eigenvectors associated with PCA are large, while numerical methods can be readily applied to the calculation of SVD.

D. Factor Analysis

While PCA tries to identify major components embedded in the data matrix, factor analysis (FA) aims to represent the data matrix and its internal relations via latent factors (variables). To do so, FA draws on a parametric model and a series of assumptions/conditions. More specifically, if words in the document-word matrix X are centered on its means in a document and we obtain a new document-word matrix X_*, we try to express the p words using latent factors:

$$Y_{n\times p} = X_*^T = F_{n\times k} A_{k\times p} + \varepsilon_{n\times p}$$

where F is a matrix containing all (latent) factors F_1, F_2,…F_k for each of n document, $A=(a_{ij})_{kXp}$ is a loading matrix representing the loadings of all words on each of the k factors, and ε is the Gaussian error term. The FA model satisfies the following four assumptions/conditions:

1. The expectation and covariance (matrix) of F are 0 and I_n, respectively;
2. The expectation and covariance (matrix) of ε are 0 and $\sigma^2_{n\times n} = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$;
3. The covariance between ε and F is 0;
4. $Cov(\mathrm{Y}) = AA^T + \sigma^2$ and $Cov(Y, F) = A_{k\times p}$.

To verify the last condition, for a centered document-word matrix X we have:

$$\begin{aligned} YY^T &= (FA+\varepsilon)(FA+\varepsilon)^T \\ &= FAA^T F^{\mathrm{T}} + \varepsilon A^T F^{\mathrm{T}} + FA\varepsilon^T + \varepsilon\varepsilon^T \end{aligned}$$

Given that $E(\varepsilon) = 0$ and $Cov(F) = I$, we can take the expectation of both sides and obtain $Cov(\mathrm{Y}) = AA^T + \sigma^2$. This conclusion has two implications. First, it is possible to calculate the loading matrix A first and then solve the latent factors using $F = \Sigma y A^T$. Second, for the i^{th} row a_i in A and a word y_i across all observations (i.e., documents), we have $\mathrm{var}(y_i) = a'_i a_i + \sigma_i^2$ and $\mathrm{cov}(y_i, y_k) = a'_i a_k$. The sum of squared loadings of y_i on all factors, or $a'_i a_i$ (i.e., the common variance), denotes the dependence of y_i on all factors, or the extent to which y_i is explained by all factors.

Factor analysis can be implemented in different ways and this study adopts the EM algorithm to conduct factor analysis [11, 12]. Yet, in existing literature the link between PCA and FA has been particularly noted [7, 13]. Related to Equation (1), we have the eigenvalues of YY^T as $\lambda_1, \lambda_2, \cdots \lambda_p$, their corresponding standardized eigenvectors as $z_{y1}, z_{y2}, \cdots z_{yp}$, and $YY^T = \sum_{i=1}^{p} \lambda_i z_{yi} z_{yi}'$ given that:

$$\begin{aligned} YY^T = \Lambda_Y &= Z_Y \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix} Z_Y^T \\ &= [z_{y1} \quad z_{y2} \quad \cdots \quad z_{yp}] \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix} \begin{pmatrix} z_{y1}' \\ z_{y2}' \\ z_{y3}' \\ z_{yp}' \end{pmatrix} \\ &= [\sqrt{\lambda_1} z_{y1} \quad \sqrt{\lambda_2} z_{y2} \quad \cdots \quad \sqrt{\lambda_p} z_{yp}] \begin{bmatrix} \sqrt{\lambda_1} z_{y1}' \\ \sqrt{\lambda_2} z_{y2}' \\ \vdots \\ \sqrt{\lambda_p} z_{yp}' \end{bmatrix} \end{aligned}$$

For the vector $[\sqrt{\lambda_1} z_{y1} \quad \sqrt{\lambda_2} z_{y2} \quad \cdots \quad \sqrt{\lambda_p} z_{yp}]$, its first m entries (where m<p) provides a possible solution to A and thus correspond to m latent factors because:

$$\begin{aligned} YY^T &\approx A\hat{A}^T + \hat{\sigma}^2 \\ &== \lambda_1 z_{y1} z_{y1}' + \lambda_2 z_{y2} z_{y2}' + \cdots + \lambda_m z_{ym} z_{ym}' + \hat{\sigma}^2 . \end{aligned}$$

Finally, it should be noted that these factors identified are often rotated to achieve maximum variance so that these independent factors can have better explanatory power.

E. Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) decomposes a matrix V into two matrices W and H and all elements of the three matrices are not negative [14]:

$$V_{n\times m} = W_{n\times r} H_{r\times m}$$

where the dimension of r is often much smaller than that of m and n. The NMF has a clear advantage over other similar algorithms in computing, interpretation and data storage. By making all elements in the three matrices non-negative, any column vector v_i in V can be expressed by a weighted sum of all column vectors in W and their corresponding weights are given by elements in the i^{th} column of H:

$$v_i = h_{1i} w_1 + w_2 h_{2i} + \cdots + w_r h_{ri} = W h_i .$$

In other words, we can learn how a whole system consists of different parts via NMF. The general idea behind NMF is also inherently related to how a whole system and its relations with different parts are perceived by human beings.

The relation between NMF and topic modeling, especially probabilistic latent semantic analysis (PLSA), has been noted [15]. For the document-word matrix X, we could define elements of W as $w_{ik} = P(topic_k)P(word_i \mid topic_k)$, elements in H as $h_{kj} = P(document_j \mid topic_k)$ and have elements x_{ij} as:

$$x_{ij} = \sum w_{ik}h_{kj} = \sum P(topic_k)P(word_i \mid topic_k)P(document_j \mid topic_k)$$

The idea is similar to that of PLSA, where a probabilistic model is used to generate topics, and words/documents are further generated based on the topic distribution.

F. Latent Dirichlet Allocation

In topic modeling, LDA provides a generative statistical model allowing for observed words and documents to be explained by latent topics that captures the similarities of words/documents [1]. For a text corpus, the generative process of LDA can be briefly summarized as follows. First, the (optimal) number of topics K needs to be specified. Second, a parameter θ_i which governs the distribution of K topics in the i^{th} document, is drawn from a Dirichlet prior distribution $D(a)$. The hyper-parameter a is a K-dimensional vector with its elements (positive real numbers) denoting the relative weights of the K topics. Third, a parameter φ_k, which governs the distribution of all V words occurring in a topic k, is drawn from another Dirichlet prior distribution $D(\beta)$. The hyper-parameter β is a V-dimensional (sparse) vector with its elements denoting the relative weights of the V words. Finally, for a word in the j^{th} location of the i^{th} document, its corresponding topic $t_{i,j}$ is drawn from a multinomial distribution $M(\theta_i)$ and the word is then generated from a multinomial distribution $M(\varphi_{t_{ij}})$. Clearly, the LDA uses probabilistic models to govern the generating processes of words and topics.

III. DATA AND MEASURES

A. Data

The text corpus used in the current study was retrieved from three major newspapers in Canada with national influence: *The Globe and Mail, (The) Toronto Star* and *National Post*. All newspaper articles published in any of the three newspapers from January 1st 1977 to June 30th 2019 are retrieved as long as they contain the word "Chinese". The data retrieval process took place from 2017 to 2019. In total, 52,317, 43,529, and 23,634 articles were retrieved from *The Globe and Mail, Toronto Star* and *National Post*, respectively. Based on lists of stop words and results from preliminary data analysis, the research team performed multiple rounds of data cleaning and compiling to remove stop words and meaningless words for topic modeling (e.g., reporters' names, street address) prior to our analysis.

B. Measures

In search of the optimal number of topics K, we compare three types of measures to assess results estimated from the five topic-modeling methods: held-out likelihood (or reconstruction loss/errors when applicable), coherence statistics, and graph-based dimensionality selection [16-19]. We calculate the held-out likelihood of fitted models using 3-fold cross validation [20]. Specifically, we split the text corpus into three parts, treat one part as a test set and the other two as training sets. We then repeat the estimation process for all three parts of the text corpus and calculate the average of the held-out likelihood. It should be noted, however, the focus of the held-out-likelihood approach is the predictive power of a specific model instead of the latent structure (e.g., topics) of the text corpus at stake.

Four measures of coherence are adopted in this study: C_v, C_npmi, C_uci, U_mass [21]. If a set of statements or terms mutually support each other, we say that this set of statements is coherent. For a specific topic, these coherence measures capture the degree of semantic similarity among words in the topic, thus allow scholars to assess whether topic modeling results represent actual semantic topics or statistical artifacts. We use the average of a coherence measure of each topic as a within-topic measure of topic coherence.

These four measures of coherence can be briefly described as follows. C_uci is probably the earliest statistic proposed to address topic coherence, which uses a (size-2) sliding window and pointwise mutual information to measure the co-occurrence probability of every word pairs in a topic. It has been suggested that C_uci provides an extrinsic measure of coherence since it pairs every single word with every other word in the topic [21]. C_npmi can be viewed as an enhanced version of C_uci because the former uses normalized pointwise mutual information (NPMI) instead of pointwise mutual information. C_v is proposed most recently and deals with indirect similarities between words, that is, some words should belong to the same topic but they rarely occur together; yet, their adjacent words should look similar. For example, suppose there are two statements "McDonald makes chicken nuggets" and "KFC serves chicken nuggets", one will probably want to put McDonald and KFC together in the same topic. The mathematical details of C_v also appears to be somewhat complicated. The use of co-occurrence counts in the calculation of the NPMI of every top word to every other top word results in a set of vectors. For every top word, there is a corresponding vector. The indirect similarity is then calculated between the vector of every top word and the sum of all other top-word vectors. Cosine distance is used as a similarity measure. Finally, based on the idea that the occurrence of every top word should be supported by every preceding top word, U_mass measures the conditional probability of weaker words given the presence of their corresponding stronger words in a topic. Different from the other three measures, U_mass is an intrinsic measure since the word list needs to be ordered and a word is compared only to its preceding and succeeding words [21]. To avoid the

calculation of the logarithm of zero, a pairwise score function of the empirical conditional log-likelihood based on smoothing counts is used.

The last measure is based on graph-based dimensionality selection. Given the very large dimensions (e.g., numbers of eigenvectors) associated with about 120,000 newspaper articles, the traditional threshold of dimensionality selection (eigenvalue as 1.0) cannot be readily applied to a big-data project. We thus relies on an automatic procedure, which maximizes a simple profile likelihood function, to search for the elbow point in a scree plot [17].

IV. Results

The three types of measures based on results from the five methods of topic modeling are presented from Figure 1 to Figure 12. For the SVD (LSA) method, it is clear that the coherence statistics, especially for the C_uci and U_mass measures, favor fewer topics (see Figure 1). This opposite conclusion holds for the measure of held-out likelihood because more topics are associated with smaller errors (see Figure 3). Yet, according to the graph-based dimensionality selection, the optimal topics number appears to be 577 (see Figure 2).

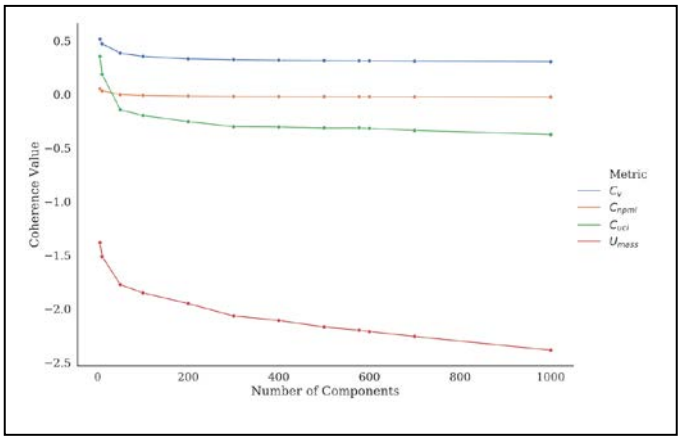
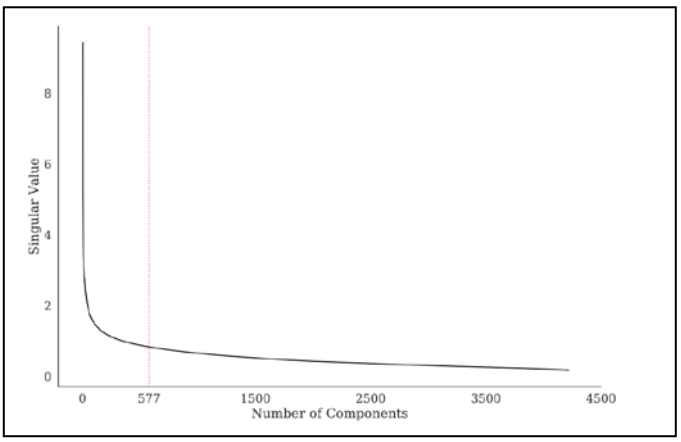


Figure 1 The SVD (LSA) method: Coherence



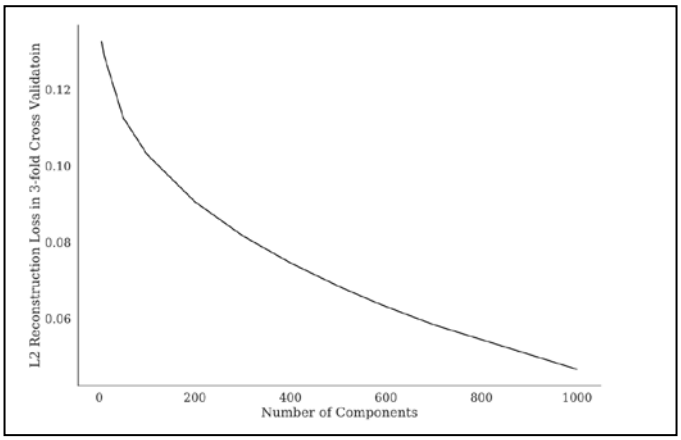


Figure 2 The SVD (LSA) method: Dimensionality selection
Figure 3 The SVD (LSA) method: Held-out error

Findings based on PCA are similar to these based on the SVD method. Coherence statistics, especially C_uci and U_mass, tend to suggest a smaller number of topics (see Figure 4). This pattern stands in contrast with the held-out likelihood, where the more the merrier (see Figure 6). The optimal number of topics suggested by dimensionality detection is 626 (see Figure 5). The coherence statistics for the FA method also prefer a smaller number of topics, although the value of U_mass slightly increases with a larger number of topics after 200 (see Figure 7). Yet, the held-out-likelihood measure of the FA model is able to specify the optimal number of topics, which appears to be 100 (see Figure 8).

The coherence statistics for the NMF methods reveal an interesting picture (see Figure 9). While the curves of C_npmi and C_v are relatively flat, results based on the C_uci and U_mass measures do not agree with each other: U_mass prefers a smaller number of topics but C_uci suggests that the value of K should be somewhere around 150 to 180. In Figure 10, the held-out error tends to support a larger number of optimal topics.

Finally, for the LDA method, the C_npmi and C_v measures do not show a strong preference over a particular number of topics (see Figure 11). The C_uci measure suggests that the value of K should be between 50 and 80 but the U_mass measure still favors a large number of topics. Finally, the held-out likelihood measure suggests that the optimal number of topics should be 20.

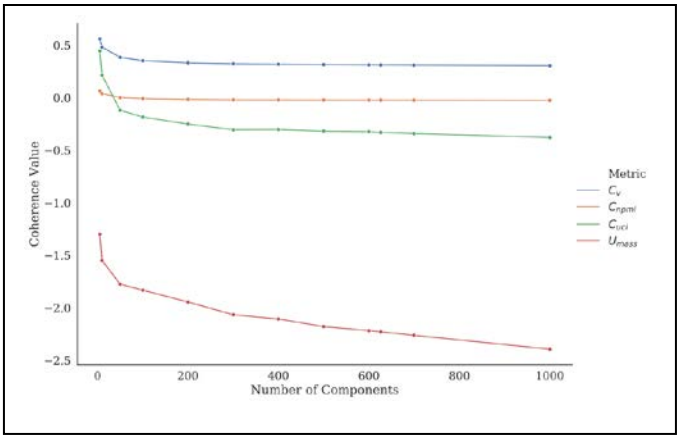


Figure 4 The PCA method: Coherence

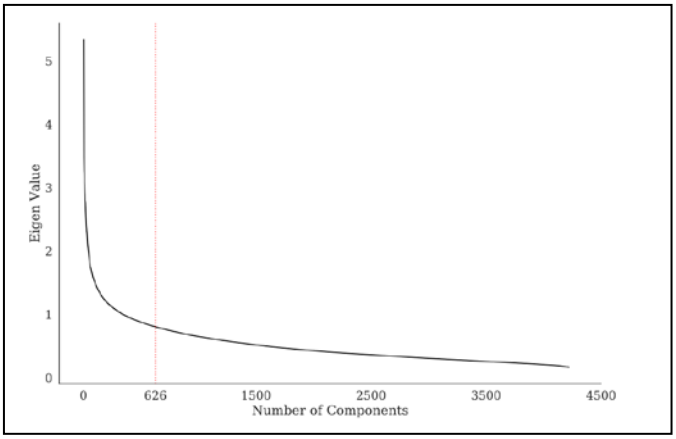


Figure 5 The PCA method: Dimensionality selection

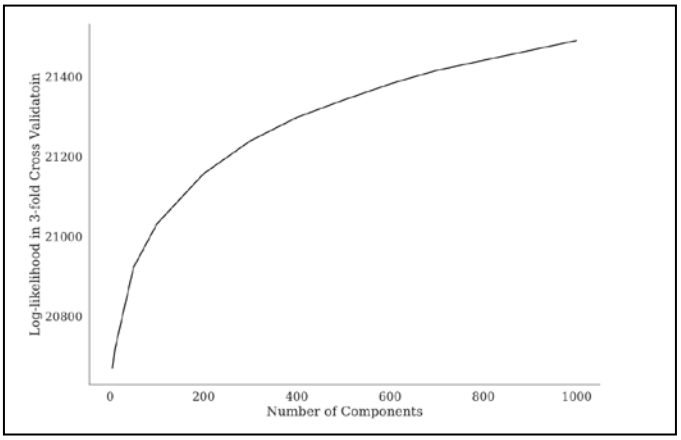


Figure 6 The PCA method: held-out likelihood

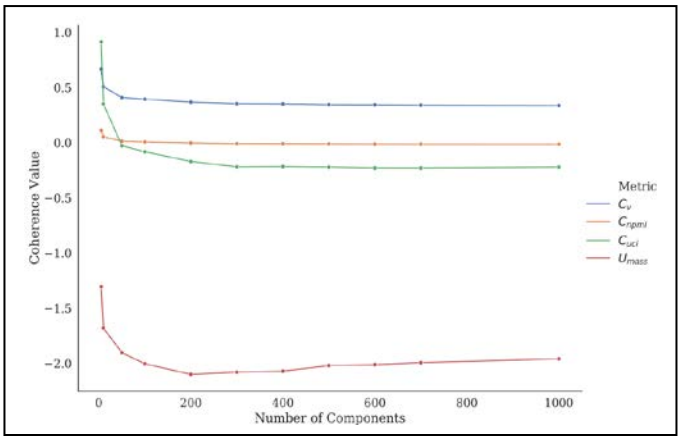


Figure 7 The FA method: Coherence

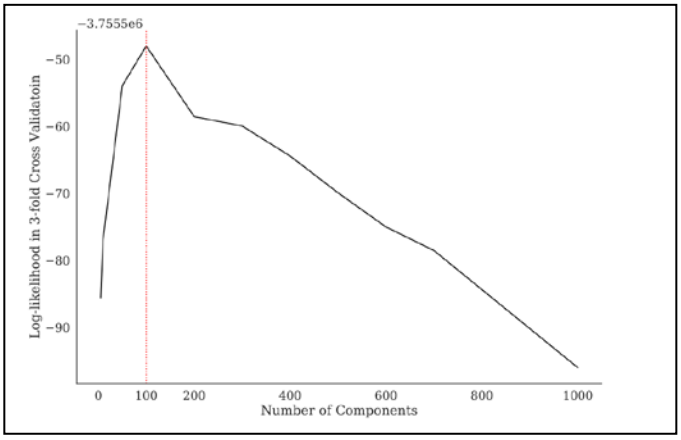


Figure 8 The FA method: held-out likelihood

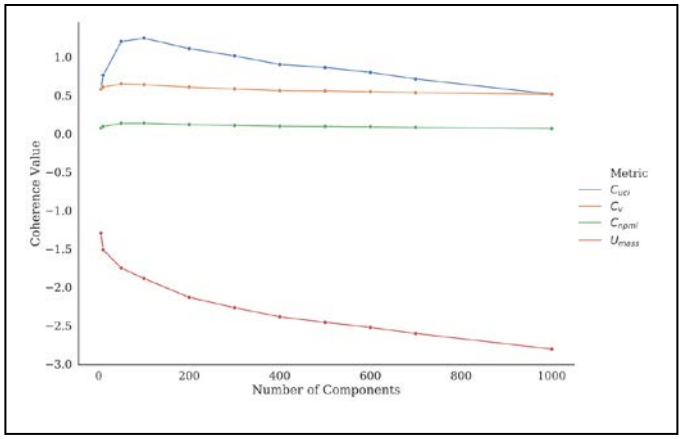


Figure 9 The NMF method: Coherence

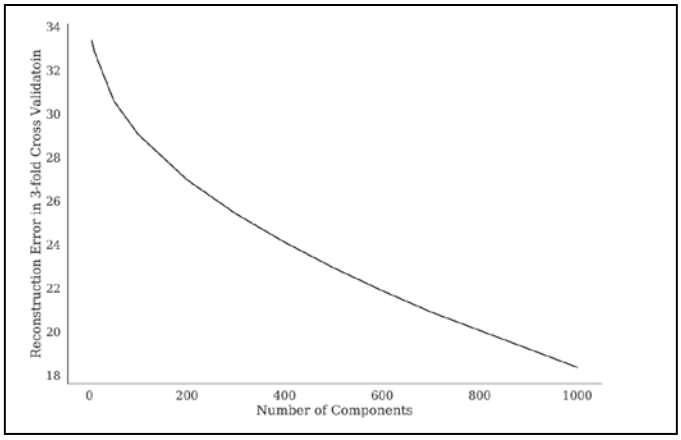


Figure 10 The NMF method: Held-out error

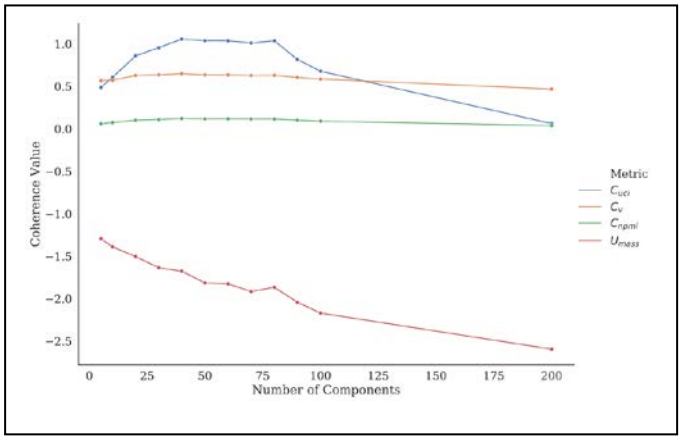


Figure 11 The LDA method: Coherence

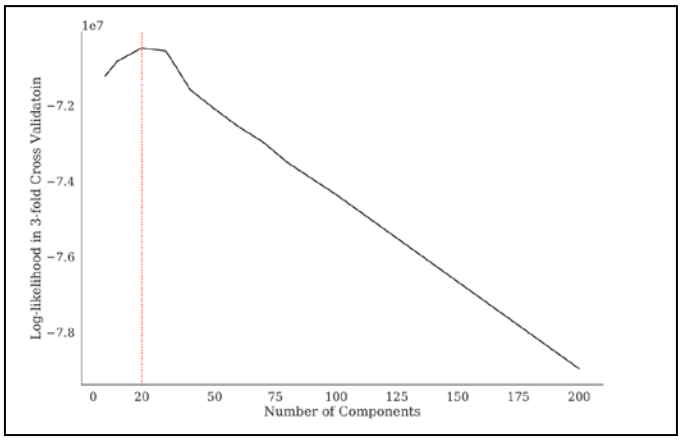


Figure 12 The LDA method: Held-out likelihood

V. Conclusion

Based on an application of five approaches to topic modeling of about 120,000 newspaper articles in Canada, major findings comparing from three measures for the optimal number of topics can be summarized in Table 1. It should be noted, however, these findings are based on a specific text corpus and can vary if other forms of data are used.

Table 1 A summary of optimal number of topics suggested by different measures and methods

	SVD	PCA	FA	NMF	LDA
C_uci	Small	Small	Small	150+	50-80
C_v	Small*	Small*	Small*	100-	25*
C_npmi	Small*	Small*	Small*	100-	25*
U_mass	Small	Small	Small	Small	Small
Held-out likelihood (error)	Large	Large	100	Large	20
Dimensionality selection	577	626	NA	NA	NA

Note: *possibly related to the scale of graphs, the conclusion suggested by this measure may not be very clear.

As suggested by Table 1, when two approaches of topic modeling are methodologically similar to each other (i.e., SVD and PCA), these measures tend to report comparable results. Yet, the optimal number of topics can vary greatly across different approaches and measures. For the same method of topic modeling, different assessment measures can also suggest different and even opposite conclusions. Among these five topic modeling methods being investigated, only assessment measures pertaining to LDA modeling tend to suggest similar numbers of optimal topics. These interesting findings beg a key question in the search of an optimal number of topics: why should measures and methods based on different methodological philosophies and computing algorithms report similar, if not identical, numbers of optimal topics? As informed by our research findings, before we ask how many optimal topics one should keep in semantic analysis, *optimal* should be first defined in terms of certain criterion, such as but not limited to, data reduction, latent structure, or predictive power.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research,* vol. 3, pp. 993-1022, 2003.

[2] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical Dirichlet processes," in *Advances in neural information processing systems*, 2005, pp. 1385-1392.

[3] D. M. Blei, "Build, compute, critique, repeat: Data analysis with latent variable models," vol. 1, pp. 203-232, 2014.

[4] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Processes: A Multidisciplinary Journal,* vol. 25, pp. 259-284, 1998.

[5] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *Journal of documentation,* vol. 60, pp. 503-520, 2004.

[6] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Information Processing Management,* vol. 39, pp. 45-65, 2003.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science,* vol. 41, pp. 391-407, 1990.

[8] J. Gao and J. Zhang, "Clustered SVD strategies in latent semantic indexing," *Information Processing Management,* vol. 41, pp. 1051-1063, 2005.

[9] G. Strang, *Introduction to linear algebra* vol. 3. Cambridge, MA: Wellesley-Cambridge Press Wellesley, MA, 1993.

[10] I. Jolliffe, *Principal component analysis*. Berlin Heidelberg: Springer, 2011.

[11] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," Technical Report CRG-TR-96-1, University of Toronto, Toronto, Canada1996.

[12] D. B. Rubin and D. T. Thayer, "EM algorithms for ML factor analysis," *Psychometrika,* vol. 47, pp. 69-76, 1982.

[13] N. Péladeau and E. Davoodi, "Comparison of latent Dirichlet modeling and factor analysis for topic extraction: A lesson of history," in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[14] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature,* vol. 401, p. 788, 1999.

[15] E. Gaussier and C. Goutte, "Relation between PLSA and NMF and implications," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, pp. 601-602.

[16] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009, pp. 288-296.

[17] M. Zhu and A. Ghodsi, "Automatic dimensionality selection from the scree plot via the use of profile likelihood," *Computational Statistics & Data Analysis,* vol. 51, pp. 918-930, 2006.

[18] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 100-108.

[19] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the conference on empirical methods in natural language processing*, 2011, pp. 262-272.

[20] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics surveys,* vol. 4, pp. 40-79, 2010.

[21] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399-408.