Finding the Number of Latent Topics With Semantic Non-Negative Matrix Factorization

RAVITEJA VANGARA, MANISH BHATTARAI, ERIK SKAU, GOPINATH CHENNUPATI, HRISTO DJIDJEV, TOM TIERNEY, JAMES P. SMITH, VALENTIN G. STANEV, AND BOIAN S. ALEXANDROV

Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
Computer, Computational and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
Amazon Alexa, Sunnyvale, CA 94089, USA
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria
Analytics, Intelligence and Technology Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
Department of Materials Science and Engineering, University of Maryland, College Park, MD 20742, USA

Corresponding author: Boian S. Alexandrov (boian@lanl.gov)

This work was supported in part by the Department of Energy (DOE) National Nuclear Security Administration (NNSA)—Office of Defense Nuclear Nonproliferation Research and Development (NA-22), in part by the LANL Laboratory Directed Research and Development (LDRD) under Grant 20190020DR and the Los Alamos National Laboratory Institutional Computing Program, and in part by the U.S. Department of Energy National Nuclear Security Administration under Contract 89233218CNA000001. The work of Hristo Djidjev was supported in part by the Science and Education for Smart Growth Operational Program (2014–2020) under Grant BG05M2OP001-1.001-0003, and in part by the European Union through the European Structural and Investment Funds.

ABSTRACT

Topic modeling, or identifying the set of topics that occur in a collection of articles, is one of the primary objectives of text mining. One of the big challenges in topic modeling is determining the correct number of topics: underestimating the number of topics results in a loss of information, i.e., omission of topics, underfitting, while overestimating leads to noisy and unexplainable topics and overfitting. In this paper, we consider a semantic-assisted non-negative matrix factorization (NMF) topics model, which we call SeNMFk, based on Kullback-Leibler(KL) divergence and integrated with a method for determining the number of latent topics. SeNMFk involves (i) creating a random ensemble of pairs of matrices whose mean is equal to the initial words-by-documents matrix representing the text corpus and the Shifted Positive Pointwise Mutual Information (SPPMI) matrix, which encodes the context information, respectively, and (ii) jointly factorizing each of these pairs with different number of topics to acquire sets of latent topics that are stable to noise. We demonstrate the performance of our method by identifying the number of topics in several benchmark text corpora, when compared to other state-of-the-art techniques. We also show that the number of document classes in the input text corpus may differ from the number of the extracted latent topics, but these classes can be retrieved by clustering the column-vectors of one of the factor matrices. Additionally, we introduce a software called pyDNMFk to estimate the number of topics. We demonstrate that our unsupervised method, SeNMFk, not only determines the correct number of topics, but also extracts topics with a high coherence and accurately classifies the documents of the corpus.

INDEX TERMS

Machine learning, NLP, topic modeling, semantic non-negative matrix factorization.

I. PRIOR PUBLICATION NOTICE

This paper is an expanded version of our work that was originally presented in the proceedings of the 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Dec 14 2020 (pp. 328-335) [DOI: 10.1109/ICMLA51294.2020.00060]. With this supplementary work, we have the opportunity to improve our prior findings, demonstrate additional use cases, and present a more detailed and accurate description of our methods. In comparison to the original publication, we establish the efficiency of our method by applying and validating SeNMFk to more benchmark datasets and providing additional information like top ten words of the extracted topics on our benchmark datasets. We introduce a new statistical measure called L-statistics that improve the automatic model determination accuracy. We expand the H-clustering technique needed to cluster accurately the documents in text corpora.

The associate editor coordinating the review of this manuscript and approving it for publication was Ikramullah Lali.
We introduce a method for SeNMFk coupled minimization based on KL-divergence and concatenation of the TF-IDF and SPPMI matrix.

- We provide a novel fast distributed Python implementation of NMF called pyDNNMFk that can be used for SeNMFk, and is back-ended for faster computation on GPUs, which is essential for analysis of large text corpora.
- Our analysis of numerical experiments includes more datasets than that of conference version.
- We introduced a new statistical measure called L-statistics that helps in automatic model determination.
- We explained in detail about the H-clustering technique that we introduced. We believe this would be a new post-processing that the topic modeling community adopts to find unique latent topics after their respective dimension reduction.
- We provided additional information in appendix like top ten words of the extracted topics on our benchmark datasets.

II. INTRODUCTION

The paradigm shifts from printed books towards proliferation of digital libraries, from printed newspapers to electronic news dissemination, and the increase in the huge repositories of web documents or hypertext from web and web-enabled applications created a colossal amount of text data. Distilling useful knowledge and keeping up with relevant information from this data is a burgeoning challenge. The extraction of relevant information and useful insights from text is called text mining [2]. Typically, text data is organized in text corpora—large collections of documents, in which each document can be modeled as a combination of recurring themes or hidden (latent) concepts called topics. These topics represent the latent (not directly observable) thematic information across a text corpus, and their extraction is known as topic modeling.

Topic modeling has been used extensively to automate the analysis of large text corpora. For example, topic modeling was utilized in social media analytics to understand conversations between people in online communities, as well as to extract useful patterns in what is shared on social media [3], [4]. It can be used to track the evolution of a software project to shed light on how the project is changing over time [5]. Topic modeling also has an immense potential in scientific research. It has been used for various task, from discovering topics in political speeches [6] to analyzing safety information in drug labels to find new uses for currently existing drugs [7]. It can even be used as a tool to accelerate the research process itself, by finding relevant publications in a constantly growing collection of scientific literature [8].

The term-frequency–inverse document frequency (TF-IDF) matrix is a heuristic document-term matrix that reflects the importance of each word in a collection of documents [9]. The latent semantic structures or topics underlying this matrix can be extracted via topic modeling, expressed as a low-rank factorization of the matrix. In the decomposition of the TF-IDF matrix, each topic consists of a collection of words, and each document is approximated by a linear combinations of these topics. As the elements of the TF-IDF matrix are strictly non-negative, the dimensionality reduction through Non-negative Matrix Factorization (NMF), where all the elements of the factor matrices are required to be non-negative, is an obvious choice for the topic modeling task.

NMF is an unsupervised learning method that approximates a non-negative matrix, $X \in \mathbb{R}^{F \times N}$, by a product of two factor matrices, $W \in \mathbb{R}^{F \times K}$, and $H \in \mathbb{R}^{K \times N}$, such that $X_{ij} \approx \sum_{k=1}^{K} W_{ik} H_{kj}$ [10]. Here $F, N$ and $K$ correspond to the number of words, the number of documents and the number of topics respectively. $W$ and $H$ are both non-negative and share a latent dimension, $k$. NMF is based on a generative statistical model, where the number of the latent features of superimposed components is the latent dimension $K$ [11]. NMF minimization is equivalent to the expectation-minimization (EM) algorithm. In this probabilistic interpretation of NMF, the observed variables are the columns $x_1, \ldots, x_N$, of the matrix, $X$, which is generated by the latent variables, $h_1, \ldots, h_K$, which are the columns of the matrix $H$. Thus, each observable $x_i$ is generated from a probability distribution with a mean $\langle x_i \rangle = \sum_{k=1}^{K} W_{ik} h_k$ [12], and the influence of $h_o$ on $x_i$ is through the basis patterns represented by the columns of the matrix $W, w_1, \ldots, w_K$. When NMF is applied to a text corpus for text mining, the basis patterns, represented by the columns of the matrix $W$, are the topics whose linear combinations span the entire corpus, and the number of topics is equal to the dimension $K$. In this case, each document is represented as a linear combination of the extracted topics with coefficients given by the corresponding columns of the matrix $H$.

Identification of the correct number of topics, called the latent dimension of a corpus, is one of the main challenges for the existing methods for topic modeling. Estimating the correct number of topics is extremely important: underestimating the number of topics results in a loss of information, i.e., omission of topics, underfitting, while overestimating leads to noisy and unexplainable topics and overfitting.

In order to address this challenge, we propose in this paper Semantic Non-negative Matrix Factorization (SeNMFk), an NMF method with a Kullback-Leibler divergence (KL) [13] foundation that incorporates: (i) the semantic structure through term-context relations with a coupled minimization of the TF-IDF matrix $X$, representing the corpus, and the Shifted Positive Point-wise Mutual Information (SPPMI) matrix, $M$, representing the semantic information; (ii) determination of the number of topics based on their stability investigated via random sampling of pairs of $X$ and $M$ matrices and a subsequent custom clustering of the sets of topic vectors with different number of topics. KL divergence measures the difference between two probability distributions by quantifying the information lost when one of the distribution is used to approximate other distribution. This method allows
us to find the latent dimension of the semantic-enhanced topics based on their stability.

Overall, the main contributions of this paper are as follows:

- We present in details our new method, SeNMFk, for topic modeling.
- We benchmark SeNMFk on nine well-known text corpora and demonstrate that SeNMFk accurately determines the number of topics in these text corpora.
- We characterize and compare the extracted latent topics with state-of-the-art topic models applied to the same datasets.
- We introduce a new statistical measure called L-statistics (the name is motivated by the L-curve) to analyze the improvements of the reconstruction error with increasing the latent dimension, which helps to identify the number of latent topics.
- We present in detail H-clustering, a procedure that clusters the coordinates of the documents in the topic space, which allows SeNMFk to accurately determine the classes of same-topic documents in the analyzed text corpora.
- We present a high performance software called pyDNMFk, to estimate the number of latent components, which scales over data of large sizes on CPU and GPU for accelerated performance.

III. RELATED WORK

A. TOPIC MODELING

The use of topic models increased due to its applicability in analyzing huge amounts of text data for exploratory analysis [14]–[16]. Some of the widely used topic modeling methods are: Latent Semantic Analysis (LSA) [17], Probabilistic LSA (PLSA) [18], Latent Dirichlet Allocation (LDA) [19], Non-negative Matrix Factorization (NMF) [20], and others.

LDA follows a generative probabilistic model, which assumes that each document can be represented as a probabilistic distribution over latent topics, and that the topics over entire corpus are shared by a common Dirichlet prior. Also, each latent topic is a probabilistic distribution over words and the distributions of words in the topics share a common Dirichlet prior [21]. LDA is successfully applied in various areas [15], [22], [23].

Another class of topic modeling methods are non-probabilistic and are based on dimension-reduction techniques, such as LSA and NMF, which decompose a matrix-based representation of the text corpus using a bag-of-words (BoW) or TF-IDF format. LSA is a factor analysis method that performs an eigenvector-based orthogonal decomposition, known as singular value decomposition (SVD), on the word-document matrix. The orthogonality constraint forces SVD to choose topics that do not overlap with one another, but leads to negative factors or components. In spite of the differences between the two modeling approaches, probabilistic and factor-based, they are closely related. NMF with KL-divergence is underpinned by a generative Poisson model [24], which is equivalent to LDA under a uniform Dirichlet prior [25], and NMF and PLSA optimize the same objective function [26].

One advantage of non-negative matrix factorization is that it provides interpretable factors, and it is widely used in text mining and document clustering [20], it also has a semi-supervised version implemented in UTOPIAN [27]. A significant limitation of the NMF approaches has been the lack of word-semantic integration that have been successfully used in deep learning as word embeddings [28], which is equivalent to factorizing the co-occurrence matrix [29].

Recently, word semantics was introduced in NMF as a regularization term based on Shifted Positive Point-wise Mutual Information (SPPMI) matrix [29]. Similarly, NMF was utilized for topic modeling of corpora of short texts [30] as well as in the Non-negative Matrix Tri-Factorization (NMTF) [31]. Such methods, which integrate NMF and word semantics, have demonstrated a better coherence of the extracted topics and high quality of the document clustering. However, they all require as a prerequisite the knowledge of the number of the latent topics.

B. MODEL DETERMINATION

The biggest challenge in topic modeling is model determination, or the identification of the accurate number of topics in the raw text corpus. The existing widely-used model determination techniques in topic modeling are as follows.

Perplexity and Coherence: These techniques use the perplexity or the coherence metrics to determine the accurate number of hidden topics. Perplexity is a heuristic measure commonly used to characterize topic quality [19], which quantifies how well a probability model predicts a sample. It has been demonstrated that perplexity does not reflect the semantic coherence of topics [32]. Therefore, coherence was introduced as a new metric [33] for estimating the quality of the extracted topics. Hierarchical Latent Tree Analysis (HLTA) [34] is another method for topic modeling, that utilize coherence metric on a hierarchy of co-occurrence patterns to identify the most coherent topics.

L-curve methods: This approach can be used in different dimensionality reduction methods or in factor analysis. It determines the optimal latent dimension by finding the elbow point of the curve showing the relative error of reconstructed data from its factors at a given latent dimension, i.e., the number of topics. NMF methods commonly try to utilize the L-curve method to identify the latent dimension $K$ [35]. The L-curve method may not always show an accurate elbow point, especially in the presence of noisy data.

Probabilistic methods: For probabilistic methods, one of the popular latent dimension identification technique is the automatic relevance determination (ARD), introduced in reference [36] and utilized for NMF in reference [37]. ARD is also used in Bayesian matrix factorization. The prediction of the latent dimensionality significantly relies on assumptions on the prior distribution, which often are difficult to estimate solely from the initial data.
Stability methods: Some of the techniques identifies topics based on the stability of the clusters of the given data. Brunet et al. [38] presented a consensus clustering based framework to determine the stability of the NMF solutions on the given data $X$ obtained from random initial conditions for each explored $K$. A term-centric stability approach has been proposed in reference [39] that investigate the agreement between the ranking of the terms of several NMF-minimizations based on the Frobenius norm.

Unfortunately, for effective topic modeling and model determination, the stability approaches have difficulties identifying the correct number of topics, although the stability metrics can be used for characterization of the quality and stability of topics.

In contrast, our method, SeNMFk [1], is semantically-assisted and is an improvement over recently developed method, NMFk, [40]–[44] for determination of the number of latent topics. NMFk has previously been applied to factorize the world’s largest cancer genomics dataset and to introduce mutational signatures [45], [46], as well as in various other fields [47]–[53]. NMFk estimates the number, $K$, of latent features based on their stability. It creates a set of random matrices whose mean is equal to the TF-IDF and SPPMI matrices, $X$ and $M$, and built by perturbing each of the elements of these matrices using random uniform noise.

2) SeNMF minimization: We explore different numbers of latent topics by running the Semantic Assisted NMF (SeNMF) algorithm for each number of topics, $k$, in an interval $[k_{\text{min}}, k_{\text{max}}]$ and for each of the $M$ random realisation pairs, $(X_q, M_q)$. SeNMF is a joint factorization done through a coupled minimization with a KL-divergence of $(X_q, M_q)$, to account for the semantic information.

3) Custom clustering: For each $k \in [k_{\text{min}}, k_{\text{max}}]$, we cluster the set of the $M$ latent topics obtained by SeNMF. To extract the latent dimension, we determine the dependence of the stability of the obtained clusters and the improvement of the reconstruction error on the latent dimension, $k$. The final latent topics, $\tilde{W}$, are the medoids of the obtained stable clusters, and let $\tilde{H}$ denote the corresponding document-membership matrix.

4) $H$-clustering: We run analysis of $\tilde{H}$ to extract the document classes/clusters in the text corpus.

In the following subsections we will describe these steps in detail. Specifically, SeNMF minimization is explained in subsection IV-A. Custom resampling, custom clustering, and $H$ clustering are explained in subsections IV-B and IV-D. We include details of $L$-Statistics in subsection IV-C.

A. SeNMF: SEMANTIC ASSISTED NMF WITH KL-DIVERGENCE

SeNMFk is solving the following optimization problem, Fig. 2,

$$\min_{W \in R^{F \times k}, H \in R^{k \times N}, G \in R^{k \times F}} D_{KL}(X || WH) + \alpha D_{KL}(M || WG),$$

where $D_{KL}(X || WH)$ is the KL-divergence of $X$ and $WH$, and $\alpha$ is a regularization parameter.

B. IV. SeNMFk

A diagram illustrating our topic modeling method, SeNMFk, is presented in Fig. 1. SeNMFk includes the following components:

1) Custom resampling: We generate an ensemble of $M$ random pairs of matrices, $[X_q, M_q]_{q=1,\ldots,M}$, with means equal to the TF-IDF and SPPMI matrices, $X$ and $M$, and built by perturbing each of the elements of these matrices using random uniform noise.
where $D_{KL}(X||\hat{X}) = \sum_{i,j} (X_{i,j} \log(\frac{X_{i,j}}{\hat{X}_{i,j}}) - X_{i,j} + \hat{X}_{i,j})$ is the generalized KL divergence between matrices $X$ and $\hat{X}$, and $\alpha$ is a regularization parameter controlling the weight of the semantic matrix decomposition.

The first term in the minimization, $D_{KL}(X||WH)$, is the error of the approximation of the TF-IDF matrix $X$ by a product of a term-topic matrix, $W$, and a topic-document matrix, $H$. The term importance in each topic is determined by the $k$ columns of $W$, where $W_{i,j}$ indicates the weight of word $i$ in topic $j$. Similarly, the topic importance in each document is determined by the $k$ rows of $H$, where $H_{i,j}$ encodes the activation of topic $i$ in document $j$. The second term in the minimization, $D_{KL}(M||WG)$, injects contextual information into the term-topic matrix, $W$. This regularization compels the columns of $W$, the topics, to represent the contextual information stored in the columns of $M$. In the optimal solution, the rows of $G$ encode the importance of each topic to represent the contextual information of each word, so $G_{i,j}$ measures importance of topic $i$ to represent the contextual information of word $j$.

We efficiently solve this problem by concatenating the TF-IDF matrix $X$ with the SPMPI matrix, $M$, since

$$D_{KL}(X||WH) + \alpha D_{KL}(M||WG) = D_{KL}([X|\alpha M]||[W|H|G]),$$

and applying our open-source software pyDNMFk (see section Software) to the concatenation, $[X|\alpha M]$.

### B. MODEL DETERMINATION

SeNMFk uses the stability of the extracted latent topics for alternative values for the latent dimension $k$ in order to determine the correct $k$. Figure 1 illustrates the workflow for SeNMFk, while Algorithm 1 represents the SeNMFk pseudocode. As a first step, SeNMFk concatenates the TF-IDF matrix $X$ and the scaled with the regularization parameter $\alpha$ PMMI context matrix $\alpha M$. This concatenation results in a new matrix, $Y = [X|\alpha M]$. Next we generate a random ensemble of $r$ matrices: $(\mathbf{Y}_1; \mathbf{Y}_2; \ldots; \mathbf{Y}_r)$ with mean value equal to $Y$. Further, we factorize each one of these matrices, $\mathbf{Y}_q$, and successively explore all possible values of the number of latent topics, $k = 1, 2, 3, \ldots, N - 1$. For each $k$, this results in a set of topics $W^k = (W^k_1; W^k_2; \ldots; W^k_r)$ and the corresponding coefficients, $H^k = ([H^kG]^N_1; [H^kG]^N_2; \ldots; [H^kG]^N_r)$. For each explored $k$, we cluster the set of columns in $W^k$, obtained via factoring each one of the $r$ perturbed versions of $Y$, as shown in Algorithm 1.

#### Algorithm 1 SeNMFk($X, M, k_{min}, k_{max}, r$)

**Require:** $X \in \mathbb{R}^{m \times n}_+, M \in \mathbb{R}^{m \times m}_+, k_{min}, k_{max}, r$

1:   for $k$ in $k_{min}$ to $k_{max}$ do
2:     for $q$ in $1$ to $r$ do
3:       $X', M' = Perturb(X, M)$ >> Creation of ensembles
4:       $W_q, H_q, G_q = SeNMF(X', M', k)$
5:   end for
6:   $W', H' = customCluster(W, H)$ >> Custom clustering of factor solutions
7:   $\hat{W} = mediansW(W')$ >> median $\hat{W}$ at $k$
8:   $\hat{H} = getFactor(X, \hat{W})$ >> activities $\hat{H}$ at $k$
9:   $e_k = getColumnWiseError(X, \hat{W}, \hat{H})$ >> Column-wise error distribution
10:  $s_k = clusterStability(W')$ >> Silhouettes for quality of clusters
11:  $e_k = reconstructErr(X, \hat{W}, \hat{H})$ >> Reconstruction error
12:  end for
13:  $k_{opt} = optRank([s_k, e_k] | k \in [k_{min}, k_{max}])$

**Ensure:** $k = k_{opt}$, $\tilde{W}[k] \in \mathbb{R}^{m \times k}_+, \tilde{H}[k] \in \mathbb{R}^{k \times n}$, $X \approx \tilde{W}[k]|\tilde{H}[k]$

We use custom clustering similar to $k$-means with cosine similarity, constrained in such a way that each cluster contains one column of each solution $W^k$. In the random ensemble, we have $r$ matrices and, correspondingly, $r$ solutions. This means that in each of the $k$ clusters (for the explored $k$) we will have $r$ elements/columns. For each $k$ we determine the stability of the clusters employing Silhouette statistics [54], which determine how compact are the clusters and how separate are they one from another. Silhouette values are between $[-1, 1]$, where $-1$ means bad clustering, while $+1$ means perfect clustering. SeNMFk determines the number of the latent topics as the maximum $k$ with average Silhouette close to one. To be able to determine which one of several $k$ with average Silhouettes close to one to choose we employ L-Statistics (for more details see, next subsection L-statistics, as well reference [42], [43]).

#### C. L-STATISTICS

For certain datasets, the plateau of the reconstruction error and Silhouette statistics may not allow to unambiguously determine the number of latent topics (see, e.g., Figure 5 for $s$-ng-2 and $s$-ng-5 datasets). For example, we can have high-values of the average Silhouette and a plateau in the reconstruction error for several different values of $k$. To make
our method more accurate, we introduce here a new statistical measure, called L-statistics (the name is motivated by the name L-curve), that helps in the identification of the number of topics in this case. In L-statistics, we use the distributions of the column reconstruction errors, $\epsilon_k = \|X_{:,i} - \tilde{W}_k \tilde{H}_k\|$, where $X_{:,i} = \tilde{W} \tilde{H}$. Thus, for each $k$, we have a distribution of column errors $\epsilon_k \rightarrow (\epsilon_1, \ldots, \epsilon_N)$, where each $\epsilon_k$ is a vector of column errors for a fixed number of topics as shown in Fig 3. The L-statistics compares the distributions of column errors corresponding to different $k$ by employing the two-sided Wilcoxon rank-sum test [55]. The Wilcoxon rank-sum test is a non-parametric test that evaluates whether two samples are taken from the same population. The Wilcoxon rank-sum test is used to check the hypothesis that data in the first sample and the data in second sample come from continuous distributions with equal medians, against the alternative that they are not. Hence, the p-value of the two-sided Wilcoxon rank-sum test allows to determine if the distributions of the column errors come from the same population or not (the smaller p-value means that there is stronger evidence in favor of the alternative hypothesis). This helps to recognize the latent dimensions: We are looking for the distribution of the column errors after which the next distributions (that correspond to a bigger $k$) fail the zero hypothesis of the Wilcoxon rank-sum test and hence the model is fitting the noise (Fig. 3). The L-statistics used in conjunction with the requirement that the minimum Silhouette statistics is greater than 0.55 is the criterion we use to determine the latent dimension, $k$. This is an empirical value which still indicates a stable cluster [1].

D. II–CLUSTERING: IDENTIFYING ACCURATELY CLASSES OF SIMILAR DOCUMENTS

In topic modeling, it is broadly accepted that the number of latent topics is equal to the number of document classes in the text corpus, that is, that there is a one-to-one correspondence between the word and document latent structures. This requirement is most likely inherited from the initial PLSA formulation, which strictly requires the number of the word latent classes to be equal to the number of document latent classes. However, in practice, this could be a very strong assumption and may not be fulfilled exactly, since the BoW and the text corpus are two different types of objects that can have their own latent structures, which (although related) may not be the exactly the same. This was previously pointed out when the bi-mixture PLSA has been introduced [56].

To account for the possibility that the latent structure of document classes is not the same as the topic latent structure, we additionally cluster the documents based on their latent coordinates, $H$, in a step we call $H$-clustering. Specifically, after identifying the latent topics, we cluster the columns of the matrix $H$ with a standard k-means algorithm. With each column of $H$ specifying a documents coordinates in latent space, $H$-clustering effectively clusters the documents. With the number of document clusters being unknown a priori, we systematically evaluate different document clusters using the standard Silhouette statistics, and use this to select the most suitable number of document clusters.

V. EVALUATION

We apply our method, SeNMFk, to nine text corpora with known numbers of latent topics and document classes, curated by domain experts. The next subsections include the description of the data, the preprocessing, and the experimental setup. These are followed by the model selection and importance of $H$–clustering for identifying the distinct latent topics in a given corpus. Finally, we present two measures that we use, the normalized mutual information (NMI) and the coherence of extracted topics, and demonstrate the consistency and quality of the extracted document classes as compared with the ground truth. Each of these evaluations also contains comparisons with other state-of-the-art topic models.

A. DATA

The nine text benchmarks, as shown in Table 1, include: (a) three datasets chosen for the topic modeling in [39], namely, $bbc$, $bbc-sport$, and guardian-2013, which belong to renowned news agencies – BBC and Guardian; (b) four variations of twenty news groups (ng20) [57], (s-ng- 2, 3, 4, 5); and (c) two variations of Reuters corpus [57], (Reuters - 4, 5). Of the variations of ng20 and Reuters, the documents of ng20 variations belong to 2, 3, 4 and 5 different classes for the twenty news groups. For Reuters dataset, largest number of classes for the respective benchmark decides the selection of different variations. For example, Reuters-4 contains all the documents of the Reuters corpus that belong to the top 4 categories. The respective number of “ground truth” latent topics are in Table 1.

B. PREPROCESSING

Preprocessing—converting raw text into a multidimensional sparse, vector space representation—is a significant step in any text mining project. We proceed in several distinct stages.
First, we remove punctuation and the common stop words in the English language from the text corpus. From that preprocessed text, we ignore the words that occur <20 times. The resulting term-document matrix is transformed into standard log-TF-IDF matrix, \(X\), with document-wise \(L_2\) normalization. This term frequency-inverse document frequency (TF-IDF) normalization weights the terms in term-document matrix by the logarithm of the inverse-relative frequency of the term in the document, i.e., if the \(i^{th}\) term occurs in \(d_i\) documents out of \(d\) total documents in the corpus, the TF-IDF normalization is the product of the frequency of the term with \(\log(d/d_i)\). We prepare the word-context co-occurrence matrix using a document length context window. The co-occurrence matrix is further modified into the SPPMI matrix, \(M\).

### C. EXPERIMENTAL SETUP

As described in the SeNMFk section, for each one of the benchmark text corpora, we create a random ensemble of matrices, \(\{Y_1; Y_2; \ldots; Y_r\}\) based on concatenation of the corresponding TF-IDF representation of the benchmark, \(X\), and a scaled SPPMI matrix, \(M\), with random resampling. Each member of this ensemble was sampled from a narrow uniform distribution with an element-wise mean equal to the concatenation of \(X\) and \(M\). For example, each element, \(y_{ij}^q\), of a member of the ensemble, \(Y^q\), is chosen as \(y_{ij}^q = Y_{ij}^q U(1 - \varepsilon, 1 + \varepsilon)\), where \(U(1 - \varepsilon, 1 + \varepsilon)\) is a random number uniformly distributed in the interval \((1 - \varepsilon, 1 + \varepsilon)\). The amplitude of the randomization, \(\varepsilon\), was varied between 7%, 5%, 3%, 1%, and 0.3%. We also explore different size of the ensemble \(Y\) in the ensemble from 20 up to 640, and if the stability of the clusters does not show bigger than 15% change from the previous set, then it is stopped. In each case, we run the SeNMF until convergence, which usually needs \(\approx 10,000\) steps. For each SeNMFk run, we used Non-Negative Double Singular Value Decomposition (NNDSDV) [58] to initialize the minimization. The regularization parameter \(\alpha\) was chosen to be one.

### D. MODEL SELECTION: FINDING THE NUMBER OF TOPICS

Our model selection procedure uses two criteria, a high average silhouette value, and a low relative reconstruction error. The silhouette and relative error values for selection of the number of latent topics for \(bbc\) and \(Reuters-4\) text corpora are shown in Figure 4. From the results, one can clearly see the calculated number of latent topics marked in rectangles. Figure 5 shows the topic determination for other benchmarks. The numbers of topics with high minima and mean silhouette values indicate stable or reliable SeNMF solutions. Lower silhouette values indicate that either the found clusters are not well separated, or are not tight clusters. The relative error, on the other hand, decreases monotonically with the number of topics. Typically the error decreases significantly from one
FIGURE 5. Silhouette characteristics and relative error values for NewsGroup20 (a-d), bbc-sport (e) and guardian-2013 (f).

E. H-CLUSTERING AND THE NUMBER OF DOCUMENT CLUSTERS

The labels or the “ground-truth” for the documents in a corpus are typically based on either manual clustering of the documents, or automatic clustering given additional metadata.
While for many corpora and many topic modeling methods this results in the number of topics matching the ‘ground truth’ number, occasionally (except in the PLSA case) the number of document clusters may not exactly match the number of latent topics [56]. Discrepancies between the number of document clusters and the number of topics can occur for many reasons. One common scenario is that a cluster of documents has significant variance in its term usage requiring multiple latent topics to adequately span this variance. This results in ancillary latent topics when compared to the number of document clusters.

Therefore, in order to make a “ground-truth” comparison for our method, we need to determine the number of document clusters from our selected factorization. To accomplish this, we cluster the columns of \( H \), the coefficients of the documents in the latent space, using k-means. The correct number of document clusters is selected through evaluating the silhouette statistics on the resulting clusters (see Section IV-D). To visualize the clusters in \( H \), we employ generalized barycentric coordinates [59] with vertices corresponding to each topic.

Figure 6 demonstrates this phenomenon. Here, we show that the extracted five latent topics in \( bbc \) correspond exactly to five document clusters. The values of the silhouette statistics for the k-means clustering of \( H \) are presented in Fig. 6(a), and the peak at \( k = 5 \) is clear. So, the \( bbc \) corpus has the same number of latent topics and document clusters. This fact is confirmed by the structure of the centroids of the five clusters as seen in Fig. 6(b), where each centroid corresponds to one of the extracted topics with very high accuracy. This is further exhibited in Fig. 6(c), where we see the documents form tight clusters around each topic.

On the other hand, in \( bbc-sport \) (Fig. 6), our method extracted \( k = 6 \) topics (see the results in Table 2). However, the clustering of \( H \) shows 5 distinct stable clusters (Fig. 6(d)) in \( X \), and we can see that the centroid of the second cluster is predominantly a combination of the second and
third topic (Fig. 6(e)). Considering the barycentric plot shown in (Fig. 6(f)), we see the same property, that is, the class corresponding to the green color is heavily used by both the second and third topics. This shows that the corresponding documents have a wide variance along one dimension and our H-clustering finds it beneficial to combine two topics to represent this class, rather than one.

We have observed similar results with Reuters-4 (Fig 8), and Reuters-5 corpora (Fig 9), where the clustering of matrix $H$ reveals the correct number of clusters.

F. RESULTS

Table 2 shows results on the identification of the correct number of latent topics. We compare our results with those presented at the state-of-the-art methods of Greene et al. [39] and consensus clustering [38]. The existing methods fail to find $k$ in most of the benchmarks except bbc, while NMFk and our SeNMFk methods find the correct $k$ in bbc, guardian-13, and all ng20 corpora. With the additional clustering of the $H$ matrix, we find the “ground-truth” number of clusters in Reuters-4,5 and bbc-sport as seen in the Column $H$-clust. Column L-statistics presents the p-values for the L-statistics test between the selected $k$ and next $k + 1$ solution.

G. NMI AND THE QUALITY OF DOCUMENT CLUSTERING

NMI, normalized mutual information [60], is one of the widely used methods to evaluate the quality of the clusters in general. Here we use this to characterize the quality of document clustering. NMI measures the mutual dependence between the predicted and the “ground-truth” labels. The closer the predicted labels match the ground truth labels, the higher the mutual dependence.

Table 3 shows the average NMI scores of seven methods for all the nine benchmarks. We observe that our methods show better NMI scores in the majority of the benchmarks. Among all the approaches, LDA performs the worst, while [61] shows better NMI for guardian-2013, and
Skmeans performs better on bbc-sport, but it does not predict the number of classes. For the remaining benchmarks, one of our methods, NMFK or SeNMFK, with H-clustering outperforms the rest. SeNMFK solutions are obtained by taking medians of the corresponding clusters of ensemble of factor matrices obtained after resampling, as mentioned clearly in earlier sections, hence they are robust and stable. We have to point out that the NMI values for Reuters-4,5 are much smaller in comparison to the other text corpora, which indicates that the document classes (extracted by all presented methods) are not ideal.

### H. $C_v$ COHESION AND QUALITY OF THE LATENT TOPICS

To characterize the quality of the extracted latent topics, we employ the coherence metric, specifically $C_v$ coherence as mentioned in [62]. $C_v$ is demonstrated to be closest type of coherence measures to the human estimation of coherence. Also, $C_v$ coherence considers the semantic information of word-word pairs, hence it does not depend on the “ground-truth” labels. This $C_v$ coherence measure is based on calculating a context vector for a word using normalized point-wise mutual information (NPMI). Table 4 shows the calculated $C_v$ coherence of LDA, Skmeans, NMFK [45] and SeNMFK. Note that, for guardian, the coherence is not shown as the original corpus is unavailable at this time, and we do not report coherence for our H-clustering method since this does not change the topic vectors. By using this metric, for majority of the analyzed text corpora, our methods show better $C_v$ coherence. The SeNMFK derived topics for the benchmarks are presented in Appendix.

### VI. SOFTWARE

pyDNMFk\(^1\) is a high performance software written at LANL by our team for performing non-negative matrix factorization.

---

1. https://github.com/lanl/pyDNMFk
on datasets in a distributed memory fashion [63], [64]. As the implementation of the SeNMFk is based on the standard NMF for decomposition, custom clustering, and L-stats for estimating \( k \), pyDNMFk is utilized for the objective of finding the number of latent topics. It scales from a laptop to clusters with numerous nodes. So far, we have tested the software with 52k+ cores and on dense datasets (>50TB) and large sparse datasets (>9EB). An illustration of pyDNMFk
VII. CONCLUSION

In this paper, we introduce an enhanced NMF algorithm called SeNMF. SeNMF is a KL divergence based semantic-assisted NMF that correctly identifies the number of topics in text corpora. This is accomplished by employing a random ensemble based on the initial TF-IDF and SPMI matrices as well as custom clustering to determine the unknown number of latent topics. The number of topics does not always coincide with the number of classes in the input text corpus. To deal with that, we use k-means clustering of the column-vectors of H in combination with Silhouette statistics, and show that the resulting algorithm correctly identifies the number of clusters in benchmark text corpora. We demonstrate the accuracy of our method on nine text corpora, and compare our results against three state-of-the-art algorithms in predicting the number of topics, NMI, and the coherence of the extracted topics.

APPENDIX: TOP TEN WORDS
See tables 5–13.

REFERENCES

[1] R. Vangara, E. Skau, G. Chennupati, H. Djidjev, T. Tierney, J. P. Smith, M. Bhattacharjee, V. G. Stanev, and B. S. Alexandrov. “Semantic non-negative matrix factorization with automatic model determination for topic modeling,” in Proc. 19th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2020, pp. 328–335.

[2] C. C. Aggarwal, Machine Learning for Text. Cham, Switzerland: Springer, 2018.

[3] L. Hong and B. D. Davison, “Empirical study of topic modeling in Twitter,” in Proc. First Workshop Social Media Anal., 2010, pp. 80–88.

[4] T. H. Nguyen and K. Shirai, “Topic modeling based sentiment analysis on social media for stock market prediction,” in Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process., 2015, pp. 1354–1364.

[5] S. W. Thomas, B. Adams, A. E. Hassan, and D. Blostein, “Modeling the evolution of topics in source code histories,” in Proc. 5th Working Conf. Mining Softw. Repositories, 2011, pp. 173–182.

[6] D. Greene and J. P. Cross, “Unveiling the political Agenda of the European parliament plenary: A topical analysis,” in Proc. ACM Web Sci. Conf., 2015, pp. 1–10.

[7] H. Bisgin, Z. Liu, R. Kelly, H. Fang, X. Xu, and W. Tong, “Investigating drug repositioning opportunities in FDA drug labels through topic modeling,” BMC Bioinf., vol. 13, no. 15, p. S6, 2012.

[8] C. Wang and D. M. Blei, “Collaborative topic modeling for recommending scientific articles,” in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2011, pp. 488–496.

[9] K. S. Jones, “A statistical interpretation of term specificity and its application in retrieval,” J. Documentation, vol. 60, no. 5, pp. 493–502, Oct. 2004.

[10] P. Paatero and U. Tapper, “Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values,” Environmetrics, vol. 5, no. 2, pp. 111–126, 1994.

[11] C. Fowlkes and A. T. Cemgil, “Nonnegative matrix factorizations as probabilistic inference in composite models,” in Proc. 17th Eur. Signal Process. Conf., Aug. 2009, pp. 1913–1917.

[12] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, no. 6755, p. 788, 1999.

[13] S. Kullback and R. A. Leibler, “On information and sufficiency,” Ann. Math. Statist., vol. 22, no. 1, pp. 79–86, 1951.

[14] H. Bisgin, Z. Liu, R. Kelly, H. Fang, X. Xu, and W. Tong, “Investigating drug repositioning opportunities in FDA drug labels through topic modeling,” BMC Bioinf., vol. 13, no. 15, p. S6, 2012.

[15] C. C. Aggarwal, Machine Learning for Text. Cham, Switzerland: Springer, 2018.
T. Liu, N. L. Zhang, and P. Chen, “Hierarchical latent tree analysis for topic detection,” in Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases, Berlin, Germany: Springer, 2014, pp. 256–272.

A. Mirzal, “Nonparametric Tikhonov regularized NMF and its application in cancer clustering,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 11, no. 6, pp. 1208–1217, Nov. 2014.

D. J. MacKay and R. M. Neal, “Automatic relevance determination for neural networks,” Cambridge Univ., Cambridge, U.K., Tech. Rep., 1994.

V. F. Y. Tan and C. Fevotte, “Automatic relevance determination in non-negative matrix factorization with the β-divergence,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 7, pp. 1592–1605, Jul. 2012.

J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, “Metagenes and molecular pattern discovery using matrix factorization,” Proc. Nat. Acad. Sci. USA, vol. 101, no. 12, pp. 4164–4169, 2004.

D. Greene, D. O’Callaghan, and P. Cunningham, “How many topics? Stability analysis for topic models,” in Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases, Berlin, Germany: Springer, 2014, pp. 498–513.

L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, “Deciphering signatures of mutational processes operative in human cancer,” Cell Rep., vol. 3, no. 1, pp. 246–259, Jan. 2013.

B. S. Alexandrov and V. V. Vesselinov, “Blind source separation for groundwater pressure analysis based on nonnegative matrix factorization,” Water Resour. Res., vol. 50, no. 9, pp. 7332–7347, Sep. 2014.

B. Nebgen, R. Vangara, M. A. Honribados-Herrera, S. Kuksova, and B. Alexandrov, “A neural network for determination of latent dimensionality in nonnegative matrix factorization,” Mach. Learn., Sci. Technol., vol. 2, Jul. 2020, Art. no. 025012.

G. Chennupati, R. Vangara, E. Skau, H. Djidjev, and B. Alexandrov, “Distributed non-negative matrix factorization with determination of the number of latent features,” J. Supercomput., vol. 76, no. 9, pp. 7458–7488, Sep. 2020.

B. S. Alexandrov, L. B. Alexandrov, F. L. Iliev, V. G. Stanev, and V. V. Vesselinov, “Source identification by non-negative matrix factorization combined with semi-supervised clustering,” U.S. Patent S10776718, Sep. 15, 2020.

L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A. L. Børresen-Dale, and S. Boyault, “Signatures of mutational processes in human cancer,” Nature, vol. 500, no. 7463, pp. 415–421, 2013.

L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. T. Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, and S. A. Islam, “The repertoire of mutational signatures in human cancer,” Nature, vol. 578, no. 7793, pp. 94–101, Feb. 2020.

C. A. Lopez, V. V. Vesselinov, S. Gnanakaran, and B. S. Alexandrov, “Unsupervised machine learning for analysis of phase separation in ternary lipid mixture,” J. Chem. Theory Comput., vol. 15, no. 11, pp. 6343–6357, Nov. 2019.

V. Stanev, V. V. Vesselinov, A. G. Kusne, G. Antoszewski, I. Takeuchi, and B. S. Alexandrov, “Unsupervised phase mapping of X-ray diffraction data by nonnegative matrix factorization integrated with custom clustering,” NPJ Comput. Mater., vol. 4, no. 1, pp. 1–10, Dec. 2018.

B. S. Alexandrov, V. G. Stanev, V. V. Vesselinov, and K. Ø. Rasmussen, “Nonnegative tensor decomposition with custom clustering for microphase separation of block copolymers,” Stat. Anal. Data Mining, ASA Data Sci. J., vol. 12, no. 4, pp. 302–310, Aug. 2019.

M. Karimi, V. Petkova, J. M. Arara, M. J. Griffin, F. W. Sellke, A. R. Bishop, B. S. Alexandrov, and A. Usheva, “Metabolomics and the pig model reveal aberrant cardiac energy metabolism in metabolic syndrome,” Sci. Rep., vol. 10, no. 1, pp. 1–11, Dec. 2020.

G. Weissman, G. Bel, A. Ben-Gal, U. Yermiyahu, B. Alexandrov, K. Ø. Rasmussen, and O. Dahan, “Increased irrigation water salinity enhances nitrate transport to deep unsaturated soil,” Vadose Zone J., vol. 19, no. 1, Jan. 2020.

R. Vangara, K. Ø. Rasmussen, D. N. Petes, G. Bel, and B. S. Alexandrov, “Identification of anomalous diffusion sources by unsupervised learning,” Phys. Rev. Res., vol. 2, no. 2, May 2020, Art. no. 023248.

B. Zhu, M. L. Poeta, M. Costantini, T. Zhang, J. Shi, S. Sentinelli, W. Zhao, V. Pompeo, M. Cardelli, B. S. Alexandrov, and B. Otlu, “The genomic and epigenomic evolutionary history of papillary renal cell carcinomas,” Nature Commun., vol. 11, no. 1, pp. 1–13, Dec. 2020.

P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., vol. 20, no. 1, pp. 53–65, 1987.

H. B. Mann and D. R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” Ann. Math. Statist., vol. 18, no. 1, pp. 50–60, Mar. 1947.

C. Shen, T. Li, and C. H. Ding, “Integrating clustering and multi-document summarization by bi-mixture probabilistic latent semantic analysis (PLSA) with sentence bases,” in Proc. 25th AAAI Conf. Artif. Intell., 2011, pp. 1–7.

D. Cai, X. He, and J. Han, “Document clustering using locality preserving indexing,” IEEE Trans. Knowl. Data Eng., vol. 17, no. 12, pp. 1624–1637, Dec. 2005.

C. Boutsidis and E. Gallopoulos, “SVD based initialization: A head start for nonnegative matrix factorization,” Pattern Recognit., vol. 41, no. 4, pp. 1350–1356, Apr. 2008.

M. S. Floater, “Generalized barycentric coordinates and applications,” Acta Numerica, vol. 24, pp. 161–214, May 2015.

A. Strehl and J. Ghosh, “Cluster ensembles—A knowledge reuse framework for combining multiple partitions,” J. Mach. Learn. Res., vol. 3, pp. 583–617, Dec. 2002.

M. Belford, B. M. Namee, and D. Greene, “Stability of topic modeling via matrix factorization,” Expert Syst. Appl., vol. 91, pp. 159–169, Jan. 2018.

M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in Proc. 8th ACM Int. Conf. Web Search Data Mining, Feb. 2015, pp. 399–408.

M. Bhattacharai, B. Nebgen, E. Skau, M. Eren, G. Chennupati, R. Vangara, H. Djidjev, J. Patchett, J. Ahrens, and B. A. Alexandrov. (2021). PyDNMF: Python Distributed Non Negative Matrix Factorization. [Online]. Available: https://github.com/lanl/pyDNMF

M. Bhattacharai, G. Chennupati, E. Skau, R. Vangara, H. Djidjev, and B. S. Alexandrov, “Distributed non-negative tensor train decomposition,” in Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC), Sep. 2020, pp. 1–10.

RAVITEJA VANGARA received the M.S. and Ph.D. degrees from the Department of Chemical and Biological Engineering, The University of New Mexico. He is currently a Postdoctoral Research Associate at the Theoretical Division, Los Alamos National Laboratory. He develops unsupervised machine learning techniques which involve graphical clustering methods, non-negative matrix and tensor factorization techniques for pattern recognition, and latent feature extraction.

MANISH BHATTARAI received the M.S. and Ph.D. degrees from the Department of Electrical and Computer Engineering, The University of New Mexico. He is currently a Postdoctoral Research Associate with the Theoretical Division, Los Alamos National Laboratory (LANL), Los Alamos, NM, USA. At LANL, he is part of the Tensor Factorizations Group, which specializes on large scale data factorization and improving the laboratory’s high-performance processing and computing abilities. He has extensively worked on developing HPC empowered ML algorithms for mining big data, such as distributed matrix and tensor factorization. His current research interests include: machine learning, computer vision, deep learning, tensor factorizations, and high performance computing.

ERIK SKAU received the B.Sc. degree in applied mathematics and physics, and the M.Sc. and Ph.D. degrees in applied mathematics from North Carolina State University, Raleigh, NC, USA. He is currently a Scientist with the Information Sciences Group, Los Alamos National Laboratory. His research expertise includes optimization techniques for matrix and tensor decompositions.
**GOPINATH CHENNUPATI** received the Ph.D. degree from the University of Limerick, Ireland. He is currently a Machine Learning Scientist at Amazon Alexa, Sunnyvale. His works include building automatic speech recognition systems and federated learning. He works on high performance computing (HPC), performance modeling, natural language processing (NLP), deep machine learning, and high performance linear algebra.

**HRISTO DJIDJIEV** received the M.Sc. degree in applied mathematics and the Ph.D. degree in computer science from Sofia University, Bulgaria. He is currently a Computer Scientist with the Information Sciences (CCS-3) Group, Los Alamos National Laboratory (LANL). Before joining LANL as a Scientist, he worked as an Assistant Professor at Rice University and as a Senior Lecturer at The University of Warwick. He is also currently a Research Adjunct Professor at Carleton University, Ottawa, Canada.

**TOM TIERNEY** received the Ph.D. degree in physics from UC Irvine, in 2002. He is currently a Senior Scientist at Los Alamos National Laboratory (LANL) and leads an applied sciences research team of over 50 technical staff. He has been with LANL, since 1997. He has over 100 publications in inertial confinement fusion, high energy density physics, radiation hydrodynamics and transport, and advanced diagnostics.

**JAMES P. SMITH** received the Ph.D. degree in physics. For two decades, he held leadership roles at Los Alamos National Laboratory (LANL), including the Group Leader, the Division Senior Scientist, and the Technical Director for the Principal Associate Directorate of Global Security. He has led many large programs, including over 100 technical staff on the DHS National Infrastructure Simulation Analysis Center delivered analytic products and represented the USA as a Science Delegate to the Quad Working Groups.

**VALENTIN G. STANEV** received the M.S. degree in physics from Sofia University, Sofia, Bulgaria, in 2000, and the Ph.D. degree in physics from Johns Hopkins University, Baltimore, MD, USA, in 2010. After a postdoctoral position at Argonne National Laboratory, he moved to the University of Maryland, College Park, where he is currently an Associate Research Scientist at the Department of Material Science and Engineering. His research interests include applications of machine learning methods to physics and material science problems, particularly in the field of superconductivity and other quantum materials.

**BOIAN S. ALEXANDROV** received the M.S. degree in theoretical physics, the Ph.D. degree in nuclear engineering, and the second Ph.D. degree in computational biophysics. He is currently a Senior Scientist at the Theoretical Division, Los Alamos National Laboratory. He is specialized in big data analytics, nonnegative matrix and tensor factorization, unsupervised learning, and latent feature extraction.

* * *

**R. Vangara et al.**: Finding Number of Latent Topics With Semantic NMF