DETECTION OF MULTIPLE DIMENSIONALITIES IN TEXTUAL DATA

Nicholas Evangelopoulos, University of North Texas, Denton, TX, USA,
Nick.Evangelopoulos@unt.edu
Triss Ashton, University of North Texas, Denton, TX, USA, Triss.Ashton@unt.edu

ABSTRACT

Dimensionality of large document collections is often detected from elbow points in eigenvalue scree plots extracted through latent semantic analysis. In the presence of multiple dimensionalities, as is frequently the case with textual data, the single change-point assumption may fail to produce accurate dimensionality estimates. In this study we propose a multiple dimensionality approach and provide a related estimation algorithm.

Keywords: text mining, latent semantic analysis, change point, log-likelihood ratio test.

INTRODUCTION

The recent proliferation of textual data, such as news stories, blogs, e-mails, SMS messages, open-ended surveys, customer comments, corporate announcements, internal company documents, medical records, reports, etc., has presented researchers in text analytics with a number of methodological challenges. Early approaches in the quantification of textual data start with the Vector Space Model, where all dictionary terms used in a collection of documents form a multi-dimensional space used to represent the documents as vectors. This kind of approach utilizes high levels of dimensionality that can involve hundreds or thousands of dimensions.

Traditionally, the high dimensionality that characterizes textual data is reduced by some technique that retains the fewest possible dimensions while discarding very little information. The Latent Semantic Analysis (LSA) family of approaches (Deerwester et al. 1990) focuses on explaining variance in term usage patterns in the document collection and extracting principal components, or latent semantic dimensions. In the context of LSA, the identification of the optimal number of retained dimensions is a challenging problem. Bradford (2008) reviews 49 studies published between 1990 and 2007, in which various data sets were reduced by latent semantic analysis. The optimal number of dimensions reported in these studies range anywhere from 6 to 1,936. In the next section we present a brief introduction to latent semantic analysis and the problem of dimensionality selection.

SELECTING THE NUMBER OF LATENT SEMANTIC DIMENSIONS

Latent Semantic Analysis

In LSA, the original term dictionary (i.e., the full list of all terms used in a collection of documents) is reduced by removing words of low information content through a stop-list. Word variants are truncated back to their root through a stemmer routine. The resulting final list of terms is used in the compilation of the collection's Vector Space Model. In matrix notation, a term-by-document ($t \times d$) frequency count matrix, $\mathbf{X}_{t,d}$, is compiled through the employment of a

dictionary of t terms in a collection of d documents. The $\mathbf{X}_{t,d}$ matrix is then decomposed using singular value decomposition (SVD) as

$$\mathbf{X}_{t,d} = \mathbf{U}_{t,r}\Sigma_{r,r}\mathbf{V}^{T}_{r,d}, \quad (1)$$

where the resulting $\mathbf{U}_{t,r}$ matrix is a term-by-factor matrix of eigenvectors of rank r, $r = \min(t, d)$, and the $\mathbf{V}^{\mathbf{T}}_{\mathbf{r,d}}$ matrix is a factor-by-document matrix of eigenvectors also of rank r. The $\mathbf{\Sigma}_{r,r}$ matrix in (1) is a diagonal matrix which contains singular values, i.e., the square roots of the eigenvalues. Once the $\mathbf{\Sigma}_{r,r}$ matrix is extracted, it can be squared to produce the common eigenvalues that characterize covariance in $\mathbf{X}_{t,d}$, due to two sets of variables, the set of terms (if documents are viewed as observations) and the set of documents (if terms are viewed as observations). Even though alternative approaches exist, most approaches attempt to determine latent semantic dimensionality through an examination of the eigenvalues. A scree plot such as the one shown in Figure 1(a) is typically used as a visual aid in depicting the k retained dimensions. However, while the single dimensionality approach may work well for numerical factor analysis problems, language data may be characterized by multiple dimensionalities. Low dimensionality can be thought of as passages that are short distances apart, focused on the same topics, and are a function of word choice. High dimensionality is associated with passages that are further apart and describe different topics using different collections of words. A corpus then consists of alternative levels of sub-topics with latent relationships that vary in distance. Thus, a meaningful generalization of the problem of identifying the optimal k is to identify a set of m alternative dimensionalities $k_1, k_2, \ldots, k_m$, such as those depicted in Figure 1(b).

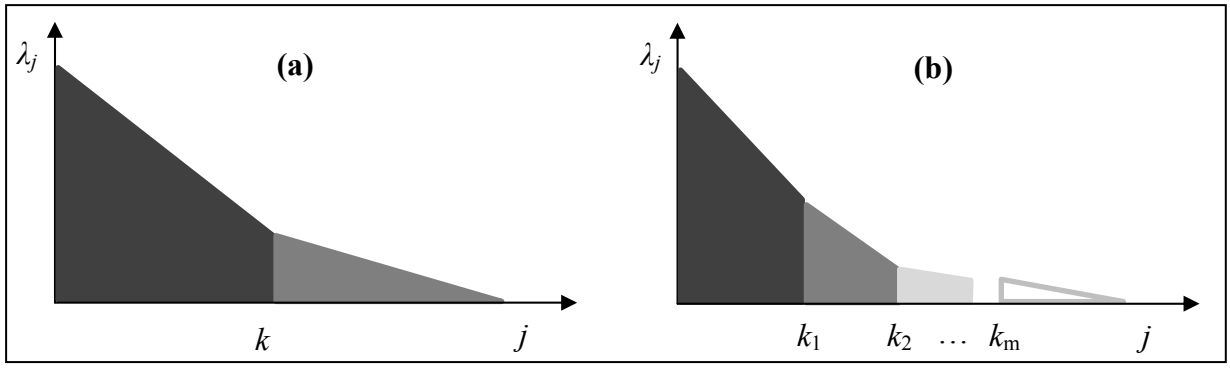

Figure 1. Single dimensionality (a) and multiple dimensionalities (b) on a scree plot

Zhu and Ghodsi (2006) introduced a method of automatically detecting the correct number of dimensions to retain. That method is based on a maximum likelihood estimator (*mle*) function that detects and estimates a single point on the scree plot that divides the set of eigenvalues into two regimes. The technique determines the last retainable eigenvalue above a "gap" or "elbow".

Dimensionality detection as a normal change point detection model

Following Zhu and Ghodsi (2006), let us assume that a sequence of ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, follow the normal distribution $F\left(\lambda; \mu_i, \sigma_i^2\right)$, $1 \leq i \leq n$. When we try to identify an "elbow point" in the scree plot, we want to test the null hypothesis of no change

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_n, \ \sigma_1 = \sigma_2 = \cdots = \sigma_n$$

against the change point alternative hypothesis

H_A: there exists an unknown k^*, $1 \le k^* \le n-1$, such that

$$\mu_1 = \cdots = \mu_{k^*} \neq \mu_{k^*+1} = \cdots = \mu_n,\ \sigma_1 = \cdots = \sigma_{k^*} \neq \sigma_{k^*+1} = \cdots = \sigma_n. \quad (2)$$

Several authors addressed the likelihood ratio method for detecting an unknown change point in time ordered data. Hinkley (1970) derived the asymptotic distribution of the *mle* of a change point in a single parameter when the parameter is both known and unknown and found the two distributions to be the same. Although the normal distribution involves more than one parameter, the basic approach derived for the single parameter case by Hinkley may also be adapted for the multi-parameter case as long as the point estimators of the parameters both before and after the change point are consistent. Gombay and Horváth (1997) studied the case of testing for the change-point in the mean of a sequence of independent random variables having a general distribution under a multidimensional setup and they also applied it to the case of a change in both parameters of a normal distribution. Following Gombay and Horváth (1997), when $k = k^*$ is known, we reject H_0 for large values of the generalized likelihood ratio

$$\Lambda_k = \frac{\sup_{\mu,\sigma} \prod_{i=1}^{k} f\left(\lambda_i; \mu, \sigma^2\right) \sup_{\mu,\sigma} \prod_{i=k+1}^{n} f\left(\lambda_i; \mu, \sigma^2\right)}{\sup_{\mu,\sigma} \prod_{i=1}^{n} f\left(\lambda_i; \mu, \sigma^2\right)}. \quad (3)$$

It has been shown in the change point literature that it is sufficient to use the maximizer of (3) as the *mle* of the unknown change point. Therefore, the point estimate for the unknown change point k^* which, in our eigenvalues application, represents the elbow point on the scree plot (Zhu and Ghodsi 2006), is

$$\hat{k}^* = \arg\max_{1 \le k \le n-1} \left(\Lambda_k\right). \quad (4)$$

Building up on (3) we derive the log-likelihood ratio

$$\begin{aligned} \log \Lambda_k = & -k \log\left(\sqrt{2\pi s_k^2}\right) - \frac{1}{2s_k^2} \sum_{i=1}^{k} \left(\lambda_i - \overline{\lambda}_k\right)^2 - (n-k) \log\left(\sqrt{2\pi s_{n-k}^{*2}}\right) \\ & - \frac{1}{2s_{n-k}^{*2}} \sum_{i=k+1}^{n} \left(\lambda_i - \overline{\lambda}_{n-k}^*\right)^2 + n \log\left(\sqrt{2\pi s_n^2}\right) + \frac{1}{2s_n^2} \sum_{i=1}^{n} \left(\lambda_i - \overline{\lambda}_n\right)^2, \end{aligned} \quad (5)$$

where all subscripts in the parameter estimates indicate the sample from which they are calculated. That is, $\overline{\lambda}_k$ is the sample mean of the first *k* eigenvalues, $\overline{\lambda}_{n-k}^*$ is the sample mean of the last $n - k$ eigenvalues, s_k^2 is the sample variance of the first *k* eigenvalues, and s_{n-k}^{*2} is the sample variance of the last $n - k$ eigenvalues. Going back to our original test of hypothesis for the single change point k^*, since k^* is unknown, we reject H_0 for large values of

$$Q_n = \max_{1 \le k \le n-1} \left(2 \log \Lambda_k\right). \quad (6)$$

The asymptotic null distribution of Q_n is given by

$$\lim_{n \to \infty} P\left\{a(\log n) Q_n^{1/2} \le x + b_2(\log n)\right\} = \exp\left(-2e^{-x}\right) \quad (7)$$

where $a(t) = (2 \log t)^{1/2}$ and $b_2(t) = 2 \log t + \log \log t$.

Note that the change point model discussed in this section is subject to the *iid* assumption on the underlying sequence of random variables. Moreover, the particular setup of the two alternatives in (2) requires that the distribution have constant vector parameter within each subsequence

(single change point assumption). This can be assessed by conducting the change-point detection test on each one of the two subsequences and failing to reject. In this paper we argue that in the presence of multiple dimensionalities, as is frequently the case with textual data, analysis of eigenvalues for detecting an elbow point on the scree plot violates the model assumption of a single change point. We address this problem by identifying a number of successive change-points in the set of eigenvalues and repeating the test on a subset of them that satisfies the single change-point assumption, producing a revised dimensionality estimate.

AN ILLUSTRATION STUDY

To illustrate the impact of multiple dimensionalities on dimensionality detection and estimation, we develop a contrived data set. In this way, we know the full structure of the data prior to analysis. We build a controlled data set using only a collection of nouns and verbs, with the understanding that most other words are excluded during the pre-processing steps which are common in LSA and other text analytic approaches. To illustrate the issues with the simple, single-dimensionality profile likelihood statistic as originally proposed by Zhu and Ghodsi (2006), our dataset consists of two dimensionality levels, expected to produce two distinct change points on the scree plot of eigenvalues. The dataset is built using a list of 60 animal names (e.g. *lions*, *tigers*, and *bears*) selected to serve as nouns and 40 action terms (e.g. *eats*, *walks*, and *flies*). The resulting dataset contains 100 terms. A representative slice of 10 sentences from the resulting data sets are presented in Table 1.

Table 1. Data used to construct contrived illustrative data set and the resulting sentences

	Noun1	Noun2	Verb1	Verb2	Resulting Sentence
1	Fox	Gorilla	begs	is shy	The Fox is shy, the Gorilla begs
2		Crocodile	discovered something		The Crocodile discovered something
3	Leopard	Ant	attacks	roams	The Ant roams, the Leopard attacks
4		Panda	attacks	roams	The Panda roams, then attacks
5	Bull	Shark	can wait	snoozes	Bull Shark can wait snoozes
6		Monkey	begs	walks	The Monkey begs while walking
7	Crab	Donkey	begs	is active	The Crab is active, the Donkey begs
8			attacks	runs	He runs, then attacks
9		Bat	cannot moo	is lazy	The Bat cannot moo because it is lazy
10	Iguana		cannot strangle you		The Iguana cannot strangle you

LSA will start by compiling a 100×150 $\mathbf{X}_{t,d}$, matrix from the illustration data set. After performing singular value decomposition on $\mathbf{X}_{t,d}$ as shown in (1), the extracted diagonal matrix $\mathbf{\Sigma}_{r,r}$ contains $r = 100$ singular values. We expect the corresponding 100 eigenvalues to exhibit three dimensional segments with two change points. The first dimensional segment should run from principal components 1 through 20 (corresponding to the 20 high frequency verbs), the second segment should run from components 21 through 40 (corresponding to the 20 lower frequency verbs), and the final segment should run from components 61 through 100 (reflecting all remaining language patterns). A scree plot of the eigenvalues extracted from the analysis of the contrived data set is shown in Figure 2. Changes in the slope are visible at approximately 20 and 40 as designed. Following Zhu and Ghodsi (2006), we analyzed the 100 extracted eigenvalues in order to detect and estimate a single elbow point on the scree plot. We obtained the test statistic for the single change point test from expression (6) as $Q_n = 163.098$ with p-value ≤ 0.0001 (obtained from expression 7). Following expression (4), the *mle* point estimate was

obtained as $\hat{k}_1^* = 39$. This result suggests that the estimator underestimated the true change point, located at $k^* = 40$. Further, the set of eigenvalues included in $[\lambda_1, \lambda_{39}]$ passes the Anderson-Darling normality test only marginally (p-value = 0.015). This result raises questions as to whether the normal Λ_k provided in (5) is really applicable. The computations reported here were performed using a custom-made Minitab macro available upon request from the corresponding author.

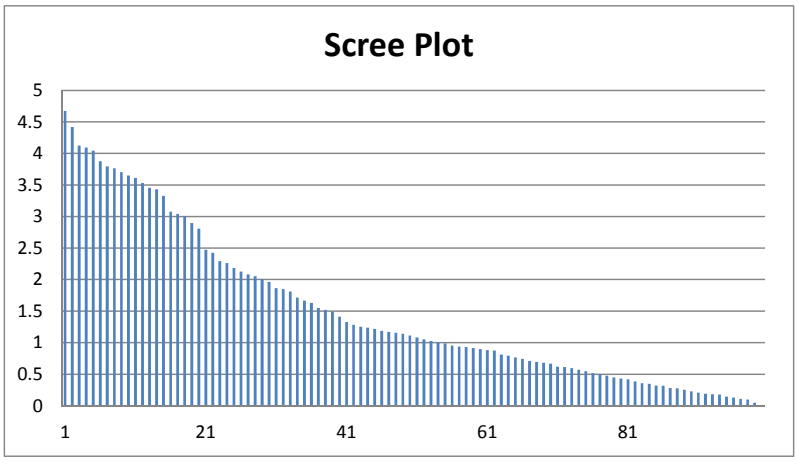

Figure 2. Scree plot of the 100 eigenvalues extracted for the contrived data set

A PROPOSED SOLUTION

What exactly went wrong with the method proposed by Zhu and Ghodsi (2006) in our illustration example presented in the previous section? In this paper we argue that *the violation of the implicit single change point assumption in the problem formulation (2) has an effect in the quality of the elbow point estimation*. We propose an approach that would iteratively examine the set of eigenvalues and, through successive segmentation, would identify segments where the eigenvalue subsets would not contain any change points. Our proposed estimation algorithm is:

1. Identify a change point that splits the set of eigenvalues in two segments
2. Iteratively process each segment by identifying change points in them, if any
3. Revise change point estimates by repeating the detection test on the reduced segments

Referring back to our analysis of the contrived data set, after testing the full set of eigenvalues for a single change point, we further test the identified segments for lack of significance, as assumed in the model formulation (2), using the asymptotic distribution in (7). However, instead of confirming lack of significance, we discover the second change point at $\hat{k}_2^* = 20$. This may explain the [1, 39] segment's failure to pass the normality test. A segmentation of the set of 100 eigenvalues into three segments results in satisfaction of the normality assumption. Corresponding Anderson-Darling (AD) goodness-of-fit tests find all three segments to be normally distributed. Those results are listed in Table 2. As a final step, after identifying the change point at $\hat{k}_2^* = 20$, and realizing that the original estimate $\hat{k}_1^* = 39$ was obtained under

violation of the single change point assumption, we perform another change point test on the segment [λ_{21}, λ_{100}]. The test produces a revised estimate for the first change point at $\hat{k}_3^* = 40$.

Table 2. Descriptive statistics and normality test results for three segments of eigenvalues

Segment	N	Mean	Standard Deviation	Minimum Value	Maximum Value	AD *p*-value	Change *p*-value
1 to 20	20	3.616	0.506	2.81	4.675	0.914	0.0119
21 to 40	20	1.919	0.319	1.4134	2.4723	0.847	0.0085
41 to 100	60	0.655	0.3803	0	1.3277	0.047	< 0.001

Concluding our discussion on the contrived data set, we note that the idea that multiple change points may exist, not only does it improve compliance with the normality assumption, if also produces a more accurate estimate. In the next section we present a small simulation study designed to explore the merit in our proposed approach of multiple elbow points.

CONCLUSION AND DIRECTIONS FOR FUTURE RESEARCH

In this paper we question the notion of single dimensionality in large unstructured document collections. We focus on the problem of identifying the right number of latent semantic dimensions through an examination of the eigenvalues and we argue that assuming single dimensionality is problematic because (1) the presence of alternative levels of semantic abstraction is expected; (2) detection of a single number of dimensions, in the presence of multiple dimensionalities, violates the assumptions of the log-likelihood ratio test; and (3) estimation under such violation may be biased. We propose an iterative to detect the presence of multiple changes in the set of eigenvalues. Applied on our illustration example, this approach produces improved estimates. Directions for future research include a description of the distribution of the dimensionality estimator and the construction of confidence intervals.

REFERENCES

Bradford, R. (2008). "An empirical study of required dimensionality for large scale latent semantic indexing applications." *In CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge mining* (New York, NY, USA, 2008), ACM, pp. 153-162.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). "Indexing by latent semantic analysis." *Journal of the American Society for Information Science,* 41, 6, pp. 391-407.

Gombay, E., and Horváth, L. (1997). An application of the likelihood method to change-point detection. *Environmetrics,* **8**, 459-467.

Hinkley, D.V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, **57**, 1-17.

Zhu, M. & Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational statistics & data analysis*, 51, pp. 918-930.