Platinum Jubilee Series

Statistical Science and
Interdisciplinary Research — Vol. 4

Advances in Multivariate Statistical Methods

Editor

Ashis SenGupta

Series Editor: ***Sankar K. Pal***

World Scientific

Chapter 7

Optimal Text Space Representation of Student Essays Using Latent Semantic Analysis

Alexander Villacorta and Sreenivasa Rao Jammalamadaka

Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA
E-mail: alex.villacorta@clearcapital.com
Email: jammalam@pstat.ucsb.edu

This paper provides a framework for optimally representing student written essays in a vector space, based upon Latent Semantic Analysis and instructor evaluated grades. Comparing student essays to an authoritative source, a ranking scheme is optimized that allows for a unique vector space representation on the unit circle. Once such a representation has been found, traditional methods of circular data analysis and inference can be applied, as we demonstrate.

Contents

7.1. Introduction

In many academic settings it is important to understand the textual writing styles of students to measure how a certain lesson makes its way into the students' writings. The focus of this article is to present a new semantic analysis technique that allows for instructors to better quantify written essays in a manner comparable to traditional numerical approaches. A typical question an instructor may be interested in is whether there is a significant amount of variability in the responses to a given question. Another question an instructor may be interested in is how deeply the course content has been retained and to know how far the student responses are, in a semantic sense, from the original source and whether this is influenced by other outside factors. Clearly, some insight into these questions is available from the grade evaluation of the teacher, although at a rough resolution. Typically, grades are given on an interval scale of $1 \ldots 10$ or on the letter grade scale of *A, B, C, D, F*. At this resolution of evaluation, inferences into the structure and variability within grade classes is difficult to assess and quantify. In addition, many educators often normalize their expectations of grades and in doing so internally adjust their grading scale, which makes comparison to other sets of essays problematic. This issue is further exaggerated when there are multiple evaluators of written text as is the case with the essay writing portion of the Standardized Aptitude Test (SAT) in the United States. To dig deeper into these questions we attempt to use current text mining approaches and circular data analysis theory to extract more specific semantic information. In this analysis, we provide a tool which can further highlight the differences in textual responses and which goes much deeper than the grades.

7.2. Textual Decomposition and Circular Analysis

Decomposition of a collection of text documents into a vector space is an active area of research with encouraging results in a wide variety of applications. In particular, a large body of effort has been dedicated to text-mining with applications to educational scoring of student essays and is best represented by the subdisciplines of eLearning and Computational Linguistics.

7.2.1. *Vector space model and latent semantic analysis*

We begin with a brief introduction to the text processing involved in setting up our method. The main vehicle for text analysis used in this work is based on the Vector Space Model (VSM). Under the VSM, each document is simply considered

to be a collection of words regardless of syntax, capitalization, or ordering of the words. In essence, the Vector Space Model creates a high dimensional document space, where each word is considered a dimension and each document lies somewhere in this space. Under this setting, document vectors may be compared just as in any other vector space. When put in matrix form with rows corresponding to unique terms and columns associated with unique documents, the matrix X_{TD} is typically called a term-document matrix. The entries (i, j) in X_{TD} can be taken to be any function of the frequency of term j in document i, such as binary indicator function, raw frequency, or the more commonly used term frequency $\times$ inverse document frequency (TFIDF) scheme , a weighting scheme which normalizes the local frequency with its global relative frequency.

The basic VSM models offer a convenient way to represent textual documents in a mathematical framework, but has several limitations in most practical applications. For instance, as more documents with disparate topics are included the size of the dictionary grows, and thus the dimensionality of the original term-document matrix also grows. Even for a modest collection of documents, the dictionary size can be very large. In addition, the distribution of words in most text collections follow a Zipf distribution implying that most words occur only once and a small set accounts for most frequency. Similarly, in most document collections there are many common words that are ambiguous or offer no semantic inference.

To solve the problems associated with the original VSM, Deerwester et al. proposed a secondary step of decomposing the term-document matrix to find a lower rank approximation to the document structure that in theory would isolate the major semantic structure of the document space. The main driving force for this is to approximate the original term document matrix with a lower order approximation based on the eigenvalues of X_{TD}. The basic algorithm of Latent Semantic Analysis (LSA) is to first construct the Term-Document Matrix X_{TD}. In the next step a Singular Value Decomposition is performed on X_{TD} such that $X_{TD} = U \times S \times V'$. The top k singular values of S are then retained and a lower rank approximation $X_{TD}^{(k)}$ is then calculated. For further explanation the reader is referred to the seminal work of Deerwester et. al. The benefits of this method are that it removes the noise associated with the collection and retains only the most prominent themes. More importantly, because of the approximation, generally the entries will not be sparse. As one can show, $X_{TD}^{(k)}$ is the best rank k model with least squares fit to X_{TD}.

In this work, we are mainly interested in measuring how far a student's writing is from the source author. Naturally, to do this we need a formal definition of what constitutes distance in the document space. For this we use the common method

of measuring distance between two document vectors by the cosine of the angle which separates them. A linear distance between two points is not used since only those texts with approximately equal number of unique terms *and* frequency would be considered close. Since we do not wish to penalize student essays that do not match the frequency of term usage, we use angular distances that better reflect similarity in concept. In n-dimensional space this takes the form of the dot product of the two vectors, when they are normalized. The cosine of the angle between two vectors, x and y, is defined as $C(x,y) = \left(\frac{x}{\|x\|}\frac{y}{\|y\|}\right)'$ and thus the angle between x and y is

$$\theta_{x,y} = \arccos(C(x,y)). \tag{7.1}$$

7.2.2. *Circular analysis*

In much of the analysis of textual collections, distances are frequently calculated among documents, with various objectives such as clustering and measuring distance distributions from a given source. As a consequence of converting a collection of documents into a vector space defined by the dictionary of words and using angles between vectors as the distance metric, the document vectors can be considered to be points on the unit hypersphere. Thus, for making inferences on such measurements, we turn to the theory of circular statistics and borrow relevant tools.

Much work has been done in the field of circular statistics, see for instance for comprehensive coverage of the field. As pointed out in these books, there is considerable difference between the treatment of linear variables and the circular case. For a set of observations on circular/angular data $\alpha_1, \ldots, \alpha_n$, a mean direction may be obtained by first treating each observation in terms of its cosine and sine components and obtaining the ***resultant vector*** as defined by

$$\mathbf{R} = (\sum_{i=1}^{n} \cos\alpha_i, \sum_{i=1}^{n} \sin\alpha_i).$$

The direction that this resultant points to, is the mean direction for the data set and at the same time, the length of this resultant vector provides a useful measure of how concentrated the data are. Define $C = \sum \cos\alpha_i$ and $S = \sum \sin\alpha_i$, then the length of the resultant vector, $|\mathbf{R}| = \sqrt{C^2 + S^2}$, is an indicator how concentrated the angles are near this mean direction. It is straightforward to see that the length of the resultant vector reaches a maximum at n and a minimum at 0. The case when $|\mathbf{R}| = 0$ corresponds to the situation where the angles are evenly distributed on the circumference and in this case a mean direction is said to not exist. Conversely, $|\mathbf{R}| = n$ when all the observations take the same value. A measure of dispersion

based of the resultant vector is given by $(n - |\mathbf{R}|)$. The direction of the resultant vector, $\hat{\alpha}_0$, is the mean direction for the circular data and is given by

$$\hat{\alpha}_0 = \arctan\left(\frac{S}{C}\right)$$

for $C > 0, S \geq 0$. For the definition of $\hat{\alpha}_0$ for other combinations of C and S see Jammalamadaka and SenGupta (2001.

Specifically, the von Mises distribution, also known as the Circular Normal, will be employed to model distances between documents in the context of analyzing student texts. The main idea behind this distribution is that the angles have unimodal distributions symmetrically distributed around a single mode, on the circle. This has many analogies to the Normal distribution on the real line. The probability distribution function for angles θ following a von Mises distribution with mean direction μ and concentration parameter κ is given below

$$f(\theta;\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(\theta-\mu)} \tag{7.2}$$

for $0 \leq \theta < 2\pi$, $0 \leq \mu < 2\pi$, and $\kappa \geq 0$ where $I_0(\kappa)$ is a modified Bessel function of the first kind. An important result shown in [Jammalamadaka and SenGupta, 2001] for large values of κ is

$$2\kappa(n - |\mathbf{R}|) \overset{approx}{\sim} \chi^2_{n-1}. \tag{7.3}$$

which follows partly because, for large values of κ, the Circular Normal distribution can be well approximated by a Normal distribution. Although the angles between two unit vectors lie in the range $[0,\pi)$, the von Mises distribution defined on the entire circle $[0,2\pi)$ provides a reasonable model for our case since a large κ ensures that the angles are tightly distributed in a small arc around the mean direction.

7.3. Optimal Vector Space Representation

One of the most critical steps in conducting a Latent Semantic Analysis on any text collection is determining the appropriate number of dimensions, k, to project the original term document matrix on. Even the authors of the seminal paper introducing LSA admit that choosing the appropriate number of dimensions to be an open research issue. In fact, determining this number is an active area of research in the information retrieval community. The reason why this poses such a challenging problem is because of the subjective nature of text. On the one hand, having too few components may lead to a compressed space which does not accurately capture the distinct semantic concepts and on the other hand retaining

too many components leads to high dimensional spaces which are known to be inefficient for measuring distances of any type. This phenomenon is known as *the curse of dimensionality*. Furthermore, in many cases a document collection may not have a strict ordering. A common example of this is seen with the results of a World Wide Web search from any search engine, such as Google. The evaluation of the ordering of the returned list is dependent on the initial objectives of the user and cannot be expected to be the same for any arbitrary user.

In the present case of analyzing student essays we are initially faced with the same challenges. However, when one is given the additional information of the instructor assessed grades for the essays, we may take this as the definitive ordering of the documents. Another way to think of this additional information is as the instructor's personal (and often internal) distance measures for each student's essay from an expected optimal essay. The assumption that an instructor's assessment in the form of grades would be available is a reasonable one since it could not be expected that student's performance on a written essay should be determined solely by a computer program.

In this work we show how to choose an optimal number of dimensions for a text space decomposition based upon a set of instructor given grades. To begin, we first discuss some score functions used to evaluate the performance of a particular text space.

7.3.1. *Loss functions*

For each lower rank approximation, $X_{TD}^{(k)}$, $k \in \{1,\ldots,min(rows(X),cols(X))\}$, a loss (or score) function is required to assess how good a match the document space is to the human judged ordering. The overall idea is that we order the semantic angular distances and partition the ordering according the the grading partitioning desired. Then, if a student is categorized in the same partition for both angular semantic distances and grades, a successful match is made. In the subsequent analysis, a document space will be created for every level of dimensionality from the full model with no dimension reduction to the overly simple case of only 1 dimension. With that in mind, we now present the following score and loss functions. For n students and $i \in \{1,\ldots,n\}$, let $d_i^{(k)}$ = angle between author text and ith student text when using a vector space model with k dimensions, where the angle is calculated as in Equation (7.1). Similarly, let y_i = grade for ith student, $y_i \in \{10,9,8,7,6,5,4,3,2,1,0\}$.

Next define Q_j, $j = 1,\ldots,J$ to be the grading partitions. For example, if three partitions are desired, a possible configuration of partitions is $Q_1 = \{10,9\}$, $Q_2 = \{8,7\}$, $Q_3 = \{6,5,4,3,2,1,0\}$, with $Q_0 \equiv \emptyset$. The introduction of Q_j is necessary

to compare the rankings of the grades and the document distances which are on different scales. To represent the students who are graded at a particular level define G_{Q_j} to be the set of all students who are graded in the jth partition, i.e.

$$G_{Q_j} = \{i : y_i \in Q_j\}.$$

Finally, we partition the ordered angular distances according to the same level as for the grades

$$T_{Q_j}^{(k)} = \{i : d_i^{(k)} \in \{d_{(|Q_{j-1}|+1)}^{(k)}, \ldots, d_{(|Q_{j-1}|+|Q_j|)}^{(k)}\}\}$$

where $d_{(m)}^{(k)}$ is the mth ordered distance for the kth document space and $|Q_j|$ denotes the cardinality of Q_j. $T_{Q_j}^{(k)}$ is simply the set of students whose angular semantic distances are partitioned in the same fashion as for their grades. Instead of using the usual form of the loss function which assigns a penalty to misclassification, we opt to instead use the equivalent score function which shows the number of correct matches so that a better intuition is gained about its results. Using this notion, we can now state the Zero-One score function for the dimension structure which produces the distances given in d as

$$S(\mathbf{d}^{(k)}, y) = \sum_{j=1}^{J} \left| G_{\{Q_j\}} \bigcap T_{\{Q_j\}}^{(k)} \right| \tag{7.4}$$

where J is the number of grading partitions. Note that for n total students the equivalent loss function is recovered as $L(d,y) = n - S(d,y)$. Figure 7.1 demonstrates the idea behind this scoring function. This function measures the amount of overlap between the assessment of the essays given by the teacher and those generated by the LSA model. However, it is of interest to understand how students interpret the material. There are various theories as to how students retain and comprehend educational material.

One point of view is that a student who understands the material will write 'close' to an authors source text. With this point of view, strong comprehension of an idea would translate to angles which are near the source document vector. Another point of view is that students comprehend the course content when they internalize the material. In this case comprehension would be represented by angles which were furthest from the source document vector, since it can be expected that a students vocabulary would be quite different from the source authors. In the following analysis we test both ideas by considering the ranking of both sets of lists.

For each set of dimensions we calculate the distances from the author as given by the angle between the vectors in the reduced principal component space. When

$Q_1 = \{10,9\}$ $Q_2 = \{8,7\}$ $Q_3 = \{6,5,4,3,2,1,0\}$

Grade	10	9	7	7	7	5	4
Student	5	2	7	1	4	3	6
	G_{Q_1}		G_{Q_2}			G_{Q_3}	
Distance	.85	.92	.96	.97	1.1	1.2	1.5
Student	4	5	3	7	1	2	6
	T_{Q_1}		T_{Q_2}			T_{Q_3}	

$$S(\mathbf{d},\mathbf{y}) = |G_{Q_1} \cap T_{Q_1}| + |G_{Q_2} \cap T_{Q_2}| + |G_{Q_3} \cap T_{Q_3}|$$
$$= 1 + 2 + 1$$
$$= 4$$

Fig. 7.1. Grading Partition Example

considering the first theory we rank these distances in ascending order from the smallest distance to the largest. In the second theory, which states that students internalize the material, distance ranking is in descending order from largest to smallest. This follows because if a student internalizes the material then they should presumably be the furthest semantically from the source author, because they use their own words.

An alternative loss function used is the more general L_1 distance. Using the notation above, suppose that there are J partitions and that for each student there is an angular distance $d_i^{(k)}$. Recall that $d_i^{(k)}$ is the angle between the source text and a student's essay when represented in a vector space with k dimensions. Then, for $T_{Q_{l_i}}^{(k)}$ under a specified ordering and a instructor assessed grade classified into $G_{Q_{m_i}}$ for the ith student, the the loss associated with these students for the given document structure is given by:

$$L(\mathbf{d}^{(k)},y) = \sum_i |l_i - m_i|, \qquad \text{for } l_i, m_i \in \{0,1,2,\ldots,J\}. \tag{7.5}$$

Equation (7.5) differs from Equation (7.4) in that it measures the severity of the misclassification. These extra penalties help to understand how different a particular angular ordering is from the grade assessments. As with the score function, the outcome from these functions are dependent on the whether ascending or descending orders are used.

For every possible vector space representation of the document collection, the two score functions described above may be applied to assess their validity, and thus identify the optimal representation of the document collection in the vector space. Once an optimal structure is found, further analysis may be based on this document vector space.

7.3.2. *Testing hypotheses using circular analysis of variance*

Through the representation of the document collection by means of the Vector Space Model, a high dimensional vector space is created where the dimensionality is determined by the vocabulary of the text corpus. By normalizing each of these vectors, one may consider that the sample of documents reside on the surface of the unit hypersphere. Further, since we are interested in how course content has disseminated among the students, we wish to investigate the pair wise distances to the source material. In doing this, the vector space is reduced to the circle since we now only consider the angles which separate a student's work from the author's.

In classical linear statistics, comparisons of the treatment effects in a population is most conveniently done using an Analysis of Variance (ANOVA) under some commonly valid assumptions. In this present case, we also wish to test hypothesis concerning the effects of various treatments and accordingly we turn to the circular analogy of the analysis of variance. One of the central assumptions of this approach is that the angular data be approximately distributed as a von Mises distribution (see Equation (7.2)). This is analogous to the linear case in that the data are approximately normally distributed. A further assumption for this approach is that the common concentration parameter κ is reasonably large. These assumptions ensure that the von Mises distribution can be well approximated by a Normal distribution, since a large value of κ means that much of the data are contained in a relatively small band of the circle. As described in Jammalamadaka and SenGupta [2001], an exact test for comparing several mean directions is not available because of the presence of the unknown nuisance parameter κ. Instead, using the assumptions that the data are distributed as a von Mises distribution and that the concentration parameter, κ, be large, an approximate test of hypothesis is used which relies on the resultant vectors.

Formally, suppose we wish to test the hypothesis that for p independent populations, the mean directions are all the same,

$$H_0 : \mu_1 = \cdots = \mu_p$$

where μ_i = mean direction for ith population. The basic intuition is that under the

Table 7.1. Approximate Circular ANOVA Table

Source of Deviation	d.f.	SS	MS	F-Stat
Between	$p-1$	$\sum_{i=1}^{p} R_i - R$	$\frac{SS_B}{p-1}$	$\frac{MS_B}{MS_W}$
Within	$n-p$	$\sum_{i=1}^{p} (n_i - R_i)$	$\frac{SS_W}{n-p}$	
Total	$n-1$	$n-R$		

null hypothesis, the direction of the resultant vector for each of the p populations should be approximately the same. Let $R = |\mathbf{R}|$ and $R_i = |\mathbf{R_i}|$ be the lengths of the result vectors for the set of angles corresponding to the combined total and for the ith population, respectively. Analogous to the linear case, $(n_i - R_i)$ can be thought of as within variation in each population, which in the circular domain translates to the within dispersion. Similarly, the total dispersion (sum of squares total in the linear case) for the data is given by $(n-R)$, where $n = \sum n_i$. It can now be shown that the total dispersion may be decomposed much like in the linear case,

$$\begin{aligned} n - R &= \left(n - \sum R_i\right) + \left(\sum R_i - R\right) \\ &= \left(\sum (n_i - R_i)\right) + \left(\sum R_i - R\right). \end{aligned} \tag{7.6}$$

By multiplying the value 2κ in Equation (7.6), we can recall the result of Equation (7.3), which leads to a similar χ^2 segmentation given by

$$\chi^2_{n-1} = \chi^2_{n-p} + \chi^2_{p-1}.$$

Then by an analogous argument to the linear case, an analysis of variance, or F-Test, may be implemented by comparing the test statistic

$$F = \frac{(\sum R_i - R)/(p-1)}{\sum (n_i - R_i)/(n-p)}$$

to the upper percentiles of an $F(\nu_1, \nu_2)$ distribution with $\nu_1 = p-1$ and $\nu_2 = n-p$. The classical ANOVA table can then be represented in the circular case as seen in Table 7.1.

It is important to mention that the arguments for an approximate analysis of variance are based on assumptions that the concentration parameter be large enough, which must be checked prior to the analysis, Jammalamadaka and Sen-Gupta show that reasonable results may be obtained when the sample mean resultant length greater than .45. When concentration assumptions cannot be met alternative techniques may be employed, for example see Jammalamadaka.

Table 7.2. Summary of case study data

	Source Authors		
	Feyerabend	Levi Strauss	Semprini
Word Count	524	634	2146
# of Essays	63	62	57
Avg Word Count	153	167	153
Italian/Non-Italian	45/18	45/17	43/14

For the context of analyzing student essays, we will employ the approximate circular analysis of variance to test for various factors and their effects on the semantic distribution.

7.4. Results

The data used for this analysis was obtained as part of a larger study conducted at the University of Basel, Switzerland by Dr. Terry Inglese and consists of student text essays from a Swiss Political Science course . Each student was required to write three essays with regard to three different philosophers: Paul Feyerabend, Claude Lévi Strauss, and Andrea Semprini. For each author, an authoritative text was available that served as the source of themes for which the students were instructed to comment on. In addition, teacher evaluated scores were also available for a total of 63 students. The data were collected throughout the entire course along with each student's native language. For simplicity we have grouped the non-native speakers into the same classification. The data are summarized in the Table 7.2 below.

To begin, we show how the students perform based on the term document matrix without any post processing but allowing for a stoplist, stemming, and TFIDF weighting scheme. It should also be noted that all of the textual essays were in Italian. Figure 7.2 shows how nearly all of the student responses to each of the three authors are near orthogonal to the original source material which is located at the due East position. The graphs were generated by mapping the high dimensional document vectors to the plane by simply graphing the angle between each essay and the target text.

These figures illustrate the idea that as the number of dimensions increase the document vector space becomes very sparse. It appears that nearly all of the textual responses are perpendicular to the source material, which illustrates the *curse of dimensionality* discussed previously.

Next we implement the LSA approach to remove the noise in the data and to try to extract the top themes in the collection. Recall that choosing the appropriate number of singular values is equivalent to choosing the appropriate number

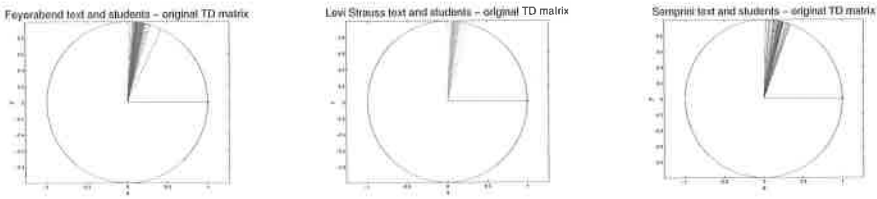


Fig. 7.2. Angular distance using original TD matrix

of principal components. One crude way of choosing an appropriate number of principal components is to investigate a scree plot, which is a graph of the ordered eigenvalues from biggest to smallest. A heuristic approach is to choose the number of components, k, such that there is minimal change in the scree plot for subsequent eigenvalues. This method is related to the percentage of variance explained in the data set by retaining the first k principal components. Figure 7.3 shows the traditional scree plot and percent of variance explained as a function of dimension. Based on the scree plot, one may consider choosing approximately 10 to 12 principal components, since those are the dimensions for which the ordered eigenvalues begin a graduated descent. Similarly, if one wished to retain a set amount of variance in the data, such as 80%, then it would be reasonable to consider using approximately 80 components. Clearly, the choice of criteria significantly affects the number of principal components retained.

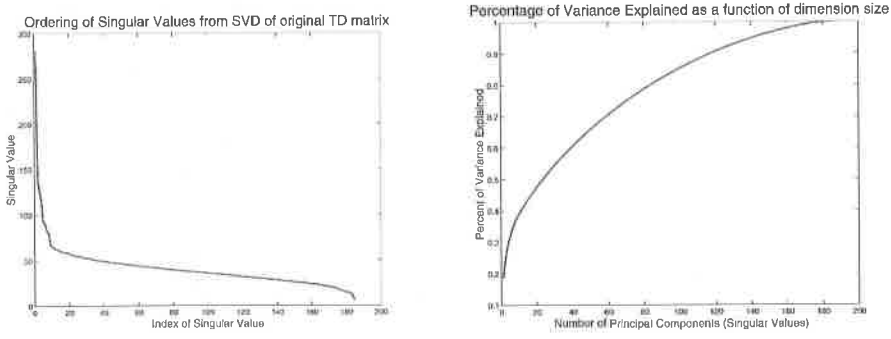
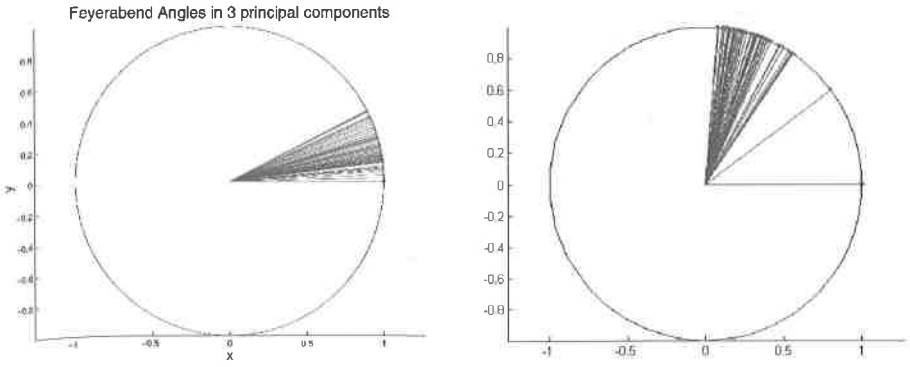


Fig. 7.3. Traditional Scree plot and Percent of Variance explained by first k principal components

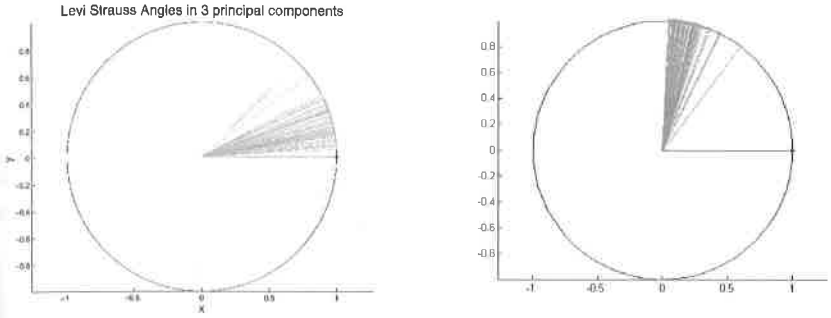
First, we show the sensitivity of angular distributions to the number of components used. Figures 7.4, 7.5, and 7.6 shows how the distribution of distances from the three authors changes for two different values of k.

More generally Figure 7.7 shows how the mean distances change as a function



(a) 3 Dims (b) 50 Dims

Fig. 7.4. Distance from Feyerabend to student essays as a function of dimension size



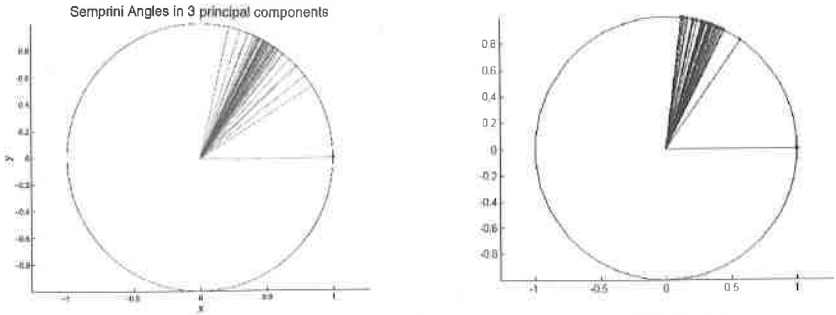
(a) 3 PCs (b) 50 PCs

Fig. 7.5. Distance from Levi Strauss to student essays as a function of dimension size

of the principal components. Clearly, the choice of dimension is critical to the conclusion of this statistical analysis. In the present work we now propose a methodology for choosing the number of principal components, which optimizes the representation of the documents with respect to the human assessed grades. Accordingly, we implement the two score functions discussed in Section 7.3.1 which help to identify when an optimal document structure is reached.

7.4.1. *Strict grade matching*

In certain instances, an instructor may be interested in knowing how varied the textual essays are within every individual grade level. Suppose that grades are



(a) 3 PCs (b) 50 PCs

Fig. 7.6. Distance from Semprini to student essays as a function of dimension size

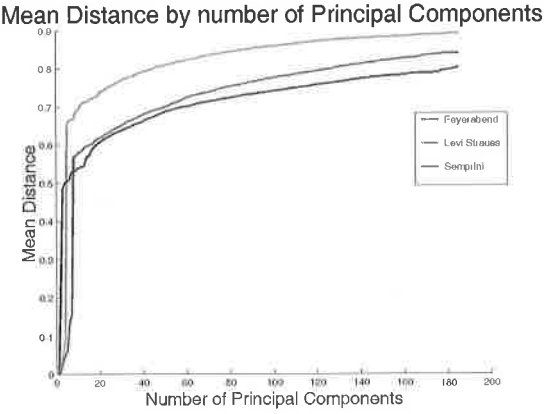


Fig. 7.7. Mean Distance from Author as a function of Principal Component

given by the instructor on an 11 point scale such that a perfect score corresponds to a grade of 10 while the score of 0 is given for incorrect answers with no partial credit. The first partitioning attempted is when the grade level resolution matches the exact assignment of grades as the teacher. That is to say when

$$Q_i = \{(10-i)\} \quad i = 0, \ldots, 10.$$

Also for each of the partitions considered we implement both ascending and descending ordering. Tables 7.3, 7.4 and 7.5 show the results of the strict partitioning, where the row 'Optimum' refers to the maximum of the Zero-One score function or the minimum of the L_1 loss function. The row 'Dim #' refers to the dimensions where the optimum value is obtained. Figure 7.8 shows a typical

Table 7.3. Summary of Loss functions for Feyerabend on 11 Grade Levels

	Feyerabend			
	$S(d,y)$		$L(d,y)$	
	A	D	A	D
Optimum	14	20	134	102
% correct	22.2%	31.8%	22.2%	20.6%
Dim #	1	167	1	8-9

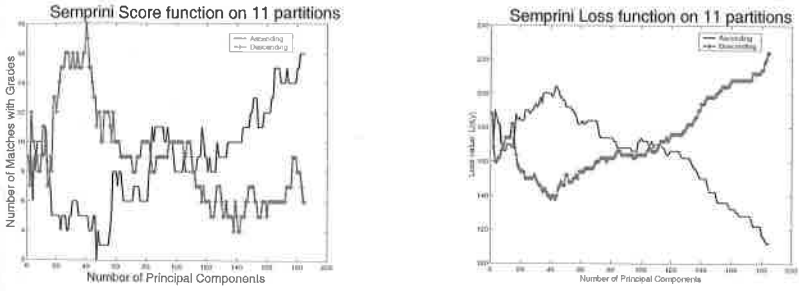
$Q_j = \{10 - j\} \quad j = 0, \ldots, 10$

Table 7.4. Summary of Loss functions for Levi Strauss on 11 Grade Levels

	Levi Strauss			
	$S(d,y)$		$L(d,y)$	
	A	D	A	D
Optimum	14	15	108	102
% correct	22.6%	24.2%	19.4%	21%
Dim #	4	5-8	180-185	55,56

$Q_j = \{10 - j\} \quad j = 0, \ldots, 10$

portrayal of the values of the score and loss function for the author Semprini, as defined in equations (7.4) and (7.5), when considering direct matches with the instructor's evaluation. As can be seen the maximum attainable match in this data set is 31.6%.



(a) Score Function (b) Loss Function

Fig. 7.8. Values of Score and Loss function for author Semprini with 11 partitions

An immediate pattern in the score and loss functions for all essays written is that the descending order better captures the grade ranking given by the instructor. When considering the 11 strict partitions, the descending order outperforms the ascending order for all three authors. For the essays written about Levi Strauss, there appears to be consistency in both the percentage of correct matches and in

Table 7.5. Summary of Loss functions for Semprini on 11 Grade Levels

	Semprini			
	$S(d,y)$		$L(d,y)$	
	A	D	A	D
Optimum	16	18	112	138
% correct	28%	31.6%	28%	31.6%
Dim #	182-185	40	183-185	40,43

$Q_j = \{10 - j\} \quad j = 0, \ldots, 10$

the dimension. However, for Feyerabend there is a significant difference between the number of matches and the dimensions where they are achieved for ascending and descending orders. The case for the essays which are written about Semprini also shows considerable differences in the optimal dimension. Furthermore, we see that the Zero-One score function consistently outperforms the L_1 loss in terms of finding the dimension with the most percentage of correct matches with the instructor grades, but this is to be expected since the L_1 loss is a generalization of the Zero-One loss. However, the benefit of using L_1 loss is that it shows other dimensions that have matches close to the Zero-One case, which in some cases be at a lower dimension number. An example of this is seen in the essays written about Feyerabend in the decreasing order setup. There we see that at compromise of approximately 11% in overlap with the instructor's grades, we may reduce the number of dimensions by over 150. Alone, this may not seem like a good trade, but when taken in conjugation with all of the other partitions, it may balance in the long run.

7.4.2. *Three level partition*

The next partition investigated is when an instructor wishes to see how the top grades are distributed. In this case, a possible ordering is to set the following partitions to

$$Q_1 = \{10,9\}, Q_2 = \{8,7\}, Q_3 = \{6,5,4,3,2,1,0\}.$$

Tables 7.6, 7.7 and 7.8 show the results of the three level partitioning. Figure 7.9 shows the values of the score and loss function for the author Levi Strauss under three level partitioning.

In this partition setup we observe that an increase in efficiency is achieved, which is to be expected since we are relaxing our restrictions on the distance ordering. However, in this case it is now observable that the descending order is best for all groups under both loss functions. This finding seems to support the case that students who understand the content better are internalizing their interpretation.

Table 7.6. Summary of Loss functions for Feyerabend on 3 Grade Levels

	Feyerabend			
	$S(d,y)$		$L(d,y)$	
	A	D	A	D
Optimum	22	33	52	36
% correct	34.9%	52.4%	34.9%	47.6%
Dim #	1,10-11	161-167	1	8,9

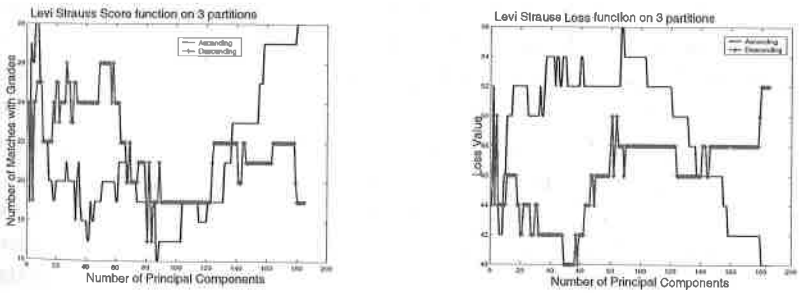
$Q_1 = \{10,9\}, Q_2 = \{8,7\}, Q_3 = \{6,5,4,3,2,1,0\}$

Table 7.7. Summary of Loss functions for Levi Strauss on 3 Grade Levels

	Levi Strauss			
	$S(d,y)$		$L(d,y)$	
	A	D	A	D
Best	28	26	40	40
%	45.1%	41.9%	45.1%	41.9%
Dim #	6-8,180-185	27,49-56,58	180-185	49-56,58

$Q_1 = \{10,9\}, Q_2 = \{8,7\}, Q_3 = \{6,5,4,3,2,1,0\}$

Also, with the exception of the essays written about the theory of Feyerabend, all of the optimal dimensions are well below the full model and seem to be less than 50 principal components. Again we observe that for the case of Feyerabend essays, a compromise of approximately 5% in overlap with the instructor's grades allows one to represent the document space with over 150 reduced dimensions.



(a) Score Function (b) Loss Function

Fig. 7.9. Values of Score and Loss function for author Levi Strauss with 3 partitions

Table 7.8. Summary of Loss functions for Semprini on 3 Grade Levels

	Semprini			
	$S(d,y)$		$L(d,y)$	
	A	D	A	D
Best	31	33	38	38
%	54.4%	57.9%	54.4%	57.9%
Dim #	182-185	40	174-185	34-36,38-40,43-44

$Q_1 = \{10,9\}, Q_2 = \{8,7\}, Q_3 = \{6,5,4,3,2,1,0\}$

Table 7.9. Summary of Loss functions for Feyerabend on 2 Grade Levels

	Feyerabend			
	$S(d,y)$		$L(d,y)$	
	A	D	A	D
Best	33	43	30	20
%	52.4%	68.3%	52.4%	68.3%
Dim #	7,36-50	2	7,36-50	2

$Q_1 = \{10,9,8,7\}, Q_2 = \{6,5,4,3,2,1,0\}$

7.4.3. *Binary partitions*

The final partition considered is when an instructor is simply interested in analyzing the scores for those students who passed the question and those who did not. In this situation, the partition is defined by

$$Q_1 = \{10,9,8,7\}, Q_2 = \{6,5,4,3,2,1,0\}.$$

Tables 9,10 and 11 show the results of the binary partitioning. An immediate observation for this case is that both the Zero-One score function and the L_1 loss function yield identical results for all three authors. Again, this is no coincidence since at a binary partition the largest difference between any two classifications can at most be one. Thus, in this case both the Zero-One and L_1 loss are equivalent. With regards to efficiency of matching the instructor's grading order, the binary partition achieves the best results, with nearly 70% accuracy in most cases. There is again a consistent gain in performance by using the descending order for distances. This further supports the idea that the student who better understands the material is internalizing the content and uses his or her own words. Finally, the dimension number which achieves the best performance is clearly less than the full model and seems to be near 40 and in the case of Feyerabend is at two dimensions.

Table 7.10. Summary of Loss functions for Levi Strauss on 2 Grade Levels

	Levi Strauss			
	$S(d,y)$		$L(d,y)$	
	A	D	A	D
Best	34	42	28	20
%	54.8%	67.7%	54.8%	67.7%
Dim #	4,106, 108-112, 126-170	44,45, 47-49, 51-64	4,106 108-112, 126-170	44,45, 47-49, 51-64

$Q_1 = \{10,9,8,7\}, Q_2 = \{6,5,4,3,2,1,0\}$

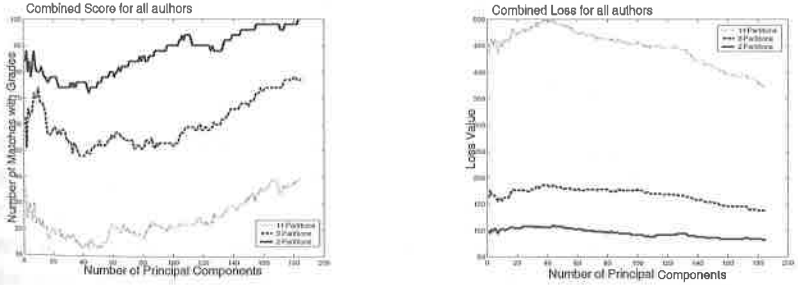
Table 7.11. Summary of Loss functions for Semprini on 2 Grade Levels

	Semprini			
	$S(d,y)$		$L(d,y)$	
	A	D	A	D
Optimum	39	39	18	18
% correct	68.4%	68.4%	68.4%	68.4%
Dim #	183-185	37-44	183-185	37-44

$Q_1 = \{10,9,8,7\}, Q_2 = \{6,5,4,3,2,1,0\}$

7.4.4. *Comparison of orderings*

After investigating various outcomes from the different partitioning levels, we now combine the scores of all three partitions to help guide a choice for an optimal dimension regardless of partitions. Figure 7.10 shows the overall distributions of the three partition levels when considering the distance ranking in ascending order. That is, when we consider that small distances relate to closer matching.

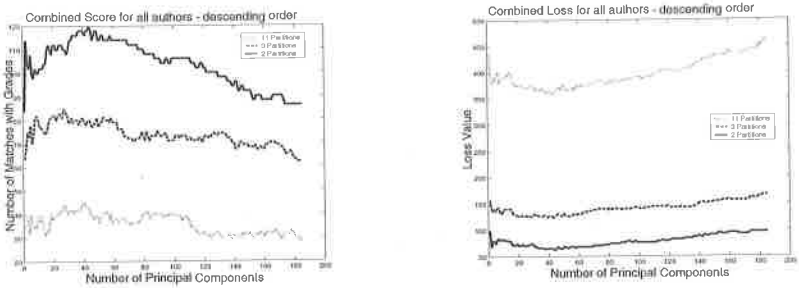


(a) Combined Score (b) Combined Loss

Fig. 7.10. Combined Score and Loss functions at all 3 partitions - Ascending Order

In the combined graph for the score function in ascending order we observe that

the optimal dimension is attained at very high levels. Specifically, the optimal dimension for *all* three partitions is at 183 to 185. This is essentially the full model with no dimension reduction. The results shown for the ascending order appear to contradict the notion that the full document space is not appropriate for semantic analysis. Similarly, the results for the L_1 loss function also show an optimal dimension at a high level. The dimension which optimizes all three grade partitions is equal to Zero-One score at levels 183 to 185.



(a) Combined Score (b) Combined Loss

Fig. 7.11. Combined Score and Loss functions at all 3 partitions - Descending Order

Figure 7.11 shows the combined scores for all authors in descending order. In this case, we see a much different story. Clearly, using descending leads to a significantly smaller dimension. The optimal dimension for all three partitions for the score function is 39 to 40. Similarly, for the L_1 loss the optimal dimension is chosen to be 43 to 44. Coupled with the efficiency results seen above, we conclude that the descending ordering appears to bear the most meaningful results in interpreting how distances should be ranked.

7.4.5. *Analysis of variance*

We next turn to an analysis of variance to demonstrate how an instructor could use the document space to make inferences on the effect of a lesson. After being satisfied with the choice of dimension for the text space we proceed to the next phase which is the analysis of the space for effects of certain factors. In the present case of student essays, we determine that descending ordering is optimal with a dimension size of 40 principal components.

Using the results introduced earlier concerning the steps required for a circular ANOVA, we begin by considering the effects of visibility of an author on students' retention of material. The two authors Feyerabend and Levi Strauss are considered

Table 7.12. Approximate Circular ANOVA Table

Source of Deviation	d.f.	SS	MS	F-Stat
Between Groups	2	.4	.2	25
Within Groups	179	1.5	.008	
Total	181	1.9		

the visible authors because of the way their lesson was interactively presented to the class. For each unit, video, images, and audio of the actual author were available for the students to explore. Conversely, the content from the author Semprini was taught in the usual textbook manner with no additional knowledge of the author. Accordingly, there are $p = 3$ populations with $n_1 = 63$, $n_2 = 62$, $n_3 = 57$, and $n = 182$. We wish to test if the mean angular distance from the author is the same for all authors. The hypothesis for this setup is

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

where μ_i is the mean direction for the ith population.
First we check to see if the estimates for the concentration parameter κ are sufficiently large, in this case they are $\hat{\kappa}_1 = 37.8$, $\hat{\kappa}_2 = 75.6$, and $\hat{\kappa}_3 = 89.8$, which are sufficiently large enough to continue with the analysis.
The sample resultant lengths are given by:

$$R_1 = 62.2$$
$$R_2 = 61.6$$
$$R_3 = 56.7$$

and the pooled sample resultant length is

$$R = 180.1.$$

Combining these results with Table 1 we obtain the following table for testing the equality of the mean directions when separated by author.

Clearly, we reject the hypothesis that angular means are all the same and upon inspection of Figure 7.7, we notice that the distances for essays written on the content of author Levi Strauss are significantly different from the other authors. Also, recall that since we accepted the descending ordering for angular distances, the essays for the theory of Levi Strauss appear to be better comprehended than the authors. This suggests that there may be validity for the positive effects of showing students pictures and videos of the authors they are learning. Of course, there are many other affecting factors that may be playing a part here, but the methodology presented above allows for further analysis of factors in a structured approach.

7.5. Conclusions

This aim of this paper is to provide a framework on how to build optimal textual vector space representations of student text essays. Using the principles demonstrated above, an instructor may investigate various hypotheses about a given data set. The key idea is that when the documents are converted to a vector space, which can be thought of as points on the hypersphere, classical linear statistics are not appropriate. In our present case, we performed a further conversion from points on the hypersphere to points on the circle since we are interested in only the distances from a common vector. Depending on the specific applications of the researcher this may not always be the correct decision.

It may be noted that for the current data set, the range of angles were limited and highly concentrated in an arc of small length, characterized by the large estimate of κ. In such a special situation, a linear ANOVA and the usual F-test can perhaps also be justified. But this is not true in general, and one should utilize the circular ANOVA, which is described here.

As a test case we investigated how information in the form of educational content is distributed in the writings of students. Traditional information retrieval approaches offer various suggestions for choosing an appropriate number of dimensions in order to represent the document space. In the work presented here, we use the grading assignment of the teacher as a score function to base the results of any given vector space model decomposition. Our results show, that students who perform well on the assignments, tend to also have writing styles that are more unique in comparison to the source material. This offers further evidence to the theory that students internalize material when they have comfortable grasp on the content.

7.6. Acknowledgements

The authors would like to sincerely thank Dr. Terry Inglese for her helpful input and supply of the test data set. This work was supported by the National Science Foundation's IGERT Program in Interactive Digital Multimedia (Award #DGE-0221713) at the University of California, Santa Barbara.

References

1. Bereiter C. and M. Scardamalia (1987), *The psychology of written composition* Lawrence Erlbaum Associates.
2. Brook Wu Y. fang and X. Chen (2005), elearning assessment through textual analy-

sis of class discussions, in Proceedings of the 5th IEEE International Conference on Advanced Learning Technologies, *ICALT*.

3. Burek G. G., M. Vargas-Vera and E. Moreale (2004), Indexing student essays paragraphs using lsa over an integrated ontological space, in COLING 2004 eLearning for Computa- tional Linguistics and Computational Linguistics for eLearning, *ed. E. H. Lothar Lem- nitzer, Detmar Meurers* (COLING, Geneva, Switzerland, August 28 2004).
4. Deerwester S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas and R. A. Harshman (1990), *Journal of the American Society of Information Science* 41, 391.
5. Haley D., P. Thomas, B. Nuseibeh, J. Taylor and P. Lefrere (2003), *E-assessment using latent semantic analysis*, in LeGE-WG 3.
6. Haley D., P. Thomas, A. D. Roeck and M. Petre (2005), A Research Taxonomy for Latent Semantic Analysis-Based Educational Applications, *Tech. Rep. TR 2005/09*, The Open University.
7. Hand D., H. Mannila and P. Smyth (2000), *Principles of Data Mining* MIT Press.
8. Mardia K.V.and P. E. Jupp(2000), *Directional Statistics* John Wiley.
9. Mayer R. (2001), *Multimedia Learning* Cambridge University Press.
10. Miller T.(2003), *Journal of Educational Computing Research* 29, 495.
11. Inglese T., R. Mayer and F. Rigotti (Feb 2007), *Learning and Instruction* 17, 67.
12. Jammalamadaka S. R.and S. Sengupta (1970), *Journal of Geology* 78, 533.
13. Jammalamadaka S. R.(1967), *Sankhya* 28, 172.
14. Jammalamadaka S. R.and A. SenGupta (2001), *Topics in Circular Statistics* World Scientic.
15. Kontostathis A., W. M. Pottenger and B. D. Davison (2005), *Identication of critical values in latent semantic indexing, in Foundations of Data Mining and Knowledge Discovery*,eds. T. Y. Lin, S. Ohsuga, C. Liau and S. Tsumoto Springer-Verlag.
16. Salton G.(1964), *A document retrieval system for man-machine interaction, in Proceedings of the 1964 19th ACM national conference*, (ACM Press, New York, NY, USA, 1964).
17. Zha H. (1998), *A Subspace-Based Model for Information Retrieval with Applications in Latent Semantic Indexing*, Tech. Rep. TR CSE-98-002, Penn State.