A Brief Survey of Latent Semantic Indexing

Edwin Montufar

PREPRINT

Latent semantic analysis is a valuable method in natural language processing for finding relationships amongst collections of documents and the words therein. Much of latent semantic analysis relies on a profound result from linear algebra concerning the singular value decomposition for general linear operators. We aim to acquaint the reader with all the subtleties underlying this important decomposition so that its application to LSI becomes apparent.

KEYWORDS: DATA MINING, INFORMATION RETRIEVAL, LSI, NLP, SVD.

1 *The Vector Space Model of Documents*

There are multiple ways to make a vector representation of a document for Latent Semantic Analysis. For our purposes, we view a document as a 'bag of words', meaning that each document is a multi-set of words. Then, the multiplicity of each element in this multi-set corresponds to the number of times a word occurs in the document. Thus, for each word in the document we tally the number of times it occurs and store its count as an entry inside in a vector, which we shall the document vector. So, for example, if we have a document, called D_i, then we denote the associated document vector as d_i, which has the underlying structure:

$$d_i = \begin{pmatrix} f_{11} \\ f_{21} \\ \vdots \\ f_{m1} \end{pmatrix}.$$

This is a column vector where each entry f_{ij}, is a non-negative integer representing the count of the term T_j occuring in document D_i $(1 \leq i \leq n, 1 \leq j \leq m)$. Now, suppose we have a collection $\mathscr{D}$ of documents $D, \ldots, D_n$, where we gather every term T_j occuring in this collection into a vocabulary set Λ. Then for every term $T_j \in \mathit{\Lambda}$, we may count the number of times it occurs for every document in $\mathscr{D}$, and store these counts inside an $m \times n$ matrix,

$$A = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & \vdots & & \vdots \\ f_{m1} & f_{m2} & \cdots & f_{mn} \end{pmatrix}. \tag{1.1}$$

Each column (row) corresponds to a document (term). Hence each entry f_{ij} is the number of times the term T_j occurs in document D_i. Now, in latent semantic analysis, we exploit a result from linear algebra stating that any $m \times n$ matrix A

with rank r can be decomposed into a product of three matrices

$$A = U\Sigma V^t \tag{1.2}$$

where the columns of U (m by m) are eigenvectors of AA^t, and the columns of V (n by n) are eigenvectors of A^tA. The r singular values on the diagonal of Σ (m by n) are the square roots of the nonzero eigenvalues of both AA^t and A^tA. This is known as the Singular Value Decomposition or SVD, which has proven very useful in many other applied settings, ranging from image compression, noise filtering, to machine learning. So, in the context of latent semantic indexing, we must find out what information the three matrices in decomposition (1.2) convey.

2 *Anatomy of the SVD Decomposition for a Term-Document Matrix*

Let the matrix U be expressed by

$$U = \begin{pmatrix} U_1 & | \, U_2 \, | & \cdots & | \, U_m \end{pmatrix} \tag{2.1}$$

where U_j are column vectors for U and $1 \leq j \leq m$. The representation of Σ is given by the block matrix

$$\Sigma = \begin{pmatrix} S & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{2.2}$$

where S is a diagonal matrix comprised of r singular values of A, denoted by σ_i. Moreover, the entries of Σ are expressed as:

$$\Sigma_{ij} = \begin{cases} \sigma_i & \text{if } i = j \leq r \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

Also, the singular values in Σ are ordered in each column such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. It is customary to represent Σ as an $r \times r$ matrix instead, ignoring all entries outside the submatrix S in (2.2). In doing so, we also decide to remove the right-most $m - r$ and right-most $n - r$ columns of U and V respectively so that

$$A = U_{m\times r}\Sigma_{r\times r}V^t_{r\times n}. \tag{2.4}$$

This form of the SVD is commonly referred to as the ***order* r *SVD*** of A, and it will be the one we work with *throughout the remaining sections*. Lastly, the matrix V^t, is represented by

Usually, in computing packages, an SVD function provides Σ as a diagonal matrix of decreasing singular values along the diagonal, omitting the columns in (1.4) that are all zero.

$$V^t = \begin{pmatrix} (V^t)_1 & | \, (V^t)_2 \, | & \cdots & | \, (V^t)_n \end{pmatrix} \tag{2.5}$$

Where $(V^t)_i$ are the column vectors for V^t and $1 \leq i \leq n$, where n corresponds to the number of documents. An important observation to make is that for matrix A, the column A_i is expressed as $A_i = U\Sigma\,(V^t)_i$, so each column $(V^t)_i$ of V^t defines the entries in the columns A_i of A. We now proceed to cover in more detail what information is conveyed entry-wise by each of these matrices .

2.1 *Semantic Dimension*

The square matrix U in (1.2) has m rows and columns corresponding to the number of terms in our entire collection of documents. With U, we focus on terms being represented as row vectors in U. Also, U is an orthogonal matrix, meaning that $UU^t = U^tU = I$; furthermore, *the columns or rows of U form an orthonormal basis for the spaces that they span*. Entries U_{ij} in U are an indication of how strongly related a term is to a topic associated with semantic dimension j. To elaborate on what we mean by semantic dimension, consider the following example: suppose that we have the following sets of terms given by

$$\mathtt{S}_1 = \{\mathtt{VECTOR,\ MATRIX,\ LINEAR,\ AUGMENTED}\},$$
$$\mathtt{S}_2 = \{\mathtt{STOCHASTIC,\ GAUSSIAN,\ DISTRIBUTION}\}.$$

Then for each term in $\mathtt{S}_1$, we may say it is related to a semantic dimension represented by the topic of linear algebra. Similarly, for $\mathtt{S}_2$, we may say each of its terms are related to a semantic dimension represented by the topic of probability. Bear in mind, that in our matrix U, there is no explicit reference to the underlying topic of each semantic dimension! Instead the topic associated with each semantic dimension is latent from the outset. Furthemore, the strengths of each semantic dimension become more apparent upon examining the singular values in Σ more closely. That being said, for the matrix Σ, we are are concerned with the magnitudes of its singular values. Thus, the magnitude of each singular value, denoted by $|\sigma_i|$, measures the importance of its corresponding semantic dimension. Essentially, larger singular values indicate that their respective semantic dimensions capture most of the variation in A. This is important because it provides us with a way reducing of the number of dimensions needed to largely preserve the informational content of the matrix A. Now, for the square matrix V^t, the entries $(V^t)_{ij}$ indicate how strongly related document D_i is to semantic dimension j. As with matrix U, V^t is *also orthogonal and its rows or columns can be used to form an orthonormal basis for the spaces that they span*.

3 *Dimensionality Reduction and Low Rank Approximation*

As hinted previously, we'll now work with the order r SVD of A, simply referring to it as the SVD of A. We saw in the previous section that the r singular values of Σ are non-increasing and that they provide a measure for how well their respective semantic dimensions capture the variation in A. That being said, those singular values Σ whose contribution to the matrix A are significant need only be retained. To understand this more precisely, we take note of a fact in linear algebra that every matrix A with rank r can be expressed as a sum of r matrices with rank 1. Hence, the SVD of A provides us with such a sum where

$$A = \sum_{i=1}^{r} U_i \sigma_i V_i^T = \sum_{i=1}^{r} \sigma_i W_i = \sum_{i=1}^{r} \langle W_i, A \rangle \, W_i. \tag{3.1}$$

Where $W_i = U_i V_i^T$ are rank 1 matrices and the singular values $\sigma_i = \langle W_i, A \rangle$ are fourier coefficients[1]. This is essentially a weighted sum, whose term-wise contributions are largely determined by the singular values. Thus, by keeping only those first few k singular values whose contribution to the variation in the sum is deemed significant, we may truncate the sum in (3.1) and obtain a lower k-rank approximation of A, denoted by A_k such that the error

$$\varepsilon = \|A - A_k\|_F = \sqrt{\sum_{i=k+1}^{r} \sigma_i^2} \tag{3.2}$$

is minimized. Indeed, from (3.2), we see that the error value is mainly determined by the truncated singular values. That being said, the lower-rank approximation of A is then determined to be

$$A_k = \sum_{i=1}^{k} U_i \sigma_i V_i^T, \tag{3.3}$$

where $k < r$. The determination of k involves a heuristic process dictated by how much of an error threshold we're willing to tolerate or how well it enables us to preserve or reveal the true informational content of A. In the latter case, the informational content of A may be contaminated with noise, therefore the truncation of singular values provides us with a way of performing noise reduction. In the context of latent semantic indexing, this is important because it provides a way of indirectly addressing ambiguity in the document collection's vocabulary or other factors in the indexing process that may contaminate the term-document matrix with noise. Also, this type of truncation may help reveal stronger connections between the semantic dimensions involved. This in a way would help provide relevant search results to a query that would otherwise go undetected. In a geometric sense, the low rank approximation of a term matrix A induces a projection of A into a lower dimensional setting where the direction of each semantic dimension becomes more constrained, in effect latent semantic connections become apparent.

3.1 A Geometric Interpretation

Suppose A is an $m \times n$ term-document matrix with rank r and that we've obtained its order r SVD. Again, let us reiterate the order r SVD of A below

$$A = U_{m\times r} \Sigma_{r\times r} V^t_{r\times n}. \tag{3.4}$$

Let us take note that the first r columns of U form an orthnormal basis $\gamma = (U_{\bullet k})_{k=1}^{r}$ for the space spanned by the columns of A, that is,

$$\operatorname{span}(\gamma) = \overbrace{\operatorname{col}(A)}^{\text{Document Space}} \tag{3.5}$$

Thus, since the columns of U constitute an orthonormal basis γ, we may represent the document vectors D_i, each as unique linear combinations of the vectors in γ,

NOTATION: The bullet in $U_{\bullet k}$ serves to indicate an aggregration over the row index, and so $U_{\bullet k}$ is representative of the k-th column in U. Similarly, $U_{1\bullet}$ is representative of the first row in U, where the position of the bullet indicates an aggregation of the column index. This notation merely serves to help us distinguish row and column vectors.

[1] In our discussion $\langle x, y \rangle$ is the standard inner product (dot product) between two vectors x, y residing in a real vector space.

moreover such linear combinations are Fourier expansions, thereby making it easy to resolve their coordinates in the space spanned by γ. Thus,

$$D_i = \sum_{k=1}^{r} \langle D_i, U_{\bullet k} \rangle \, U_{\bullet k} \quad 1 \le i \le n. \tag{3.6}$$

In addition, the n columns of the matrix $\Sigma V^t_{r\times n}$ are the coordinates of the document vectors in the space spanned by γ, which are obtained from (3.6). Hence,

$$\Sigma V^t_{r\times n} = \left([D_1]_\gamma \mid \quad [D_2]_\gamma \mid \quad \cdots \quad \mid [D_n]_\gamma \right), \quad [D_i]_\gamma = \begin{pmatrix} \langle D_i, U_{\bullet 1} \rangle \\ \langle D_i, U_{\bullet 2} \rangle \\ \vdots \\ \langle D_i, U_{\bullet r} \rangle \end{pmatrix}, \tag{3.7}$$

for $1 \le i \le n$. Now, moving on to V^t, we note that the first r rows of V^t (columns of V) constitute an orthonormal basis $\beta = (V_{k\bullet})_{k=1}^{r}$ for the space spanned by the rows of A, that is,

$$\operatorname{span}(\beta) = \overbrace{\operatorname{row}(A)}^{\text{Term Space}} = \operatorname{col}\left(A^t\right). \tag{3.8}$$

We may express each term vector as a fourier expansion over the orthonormal basis β, so that

$$T_j = \sum_{k=1}^{r} \left\langle T_j, V_{k\bullet} \right\rangle V_{k\bullet}, \quad 1 \le j \le m. \tag{3.9}$$

Moreover, the m rows of matrix $U\Sigma_{m\times r}$ comprise the coordinates of the term vectors with respect to β, that is, their coordinates in the space spanned by β. Hence,

$$U\Sigma_{m\times r} = \begin{pmatrix} [T_1]^t_\beta \\ [T_2]^t_\beta \\ \vdots \\ [T_m]^t_\beta \end{pmatrix} \quad \text{and} \quad \left[T_j\right]_\beta = \begin{pmatrix} \left\langle T_j, V_{1\bullet} \right\rangle \\ \left\langle T_j, V_{2\bullet} \right\rangle \\ \vdots \\ \left\langle T_j, V_{r\bullet} \right\rangle \end{pmatrix}, \quad 1 \le j \le m. \tag{3.10}$$

Now, the bases γ and β define orthornormal axes for the r dimensional spaces that they span (i.e. document and term space, respectively). We can think of each axis as respresenting a "semantic dimension". So, by looking at the r singular values along the diagonal of Σ, we are essentially looking at the scaling factors of each axes, which in effect measure the relative importance of each semantic dimension. Since the r singular values of Σ satisfy $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r$, it should be no surprise if the first semantic dimension is of greater importance. To see how this scaling is done explicitly for the orthonormal axes of the document space, we again recall that the basis γ of the document space is comprised of the columns of U , so all we need to do is examine the matrix $U\Sigma_{m\times r}$, and observe that its r columns are of the form

$$(U\Sigma)_{\bullet k} = \begin{pmatrix} U_{1k}\sigma_k \\ U_{2k}\sigma_k \\ \vdots \\ U_{mk}\sigma_k \end{pmatrix} = \sigma_k \begin{pmatrix} U_{1k} \\ U_{2k} \\ \vdots \\ U_{mk} \end{pmatrix}, \quad k = 1, 2, \ldots, r. \tag{3.11}$$

Hence, each k-th column of U is weighted by the k-th singular value which is a scalar, thereby scaling the axes of the document space. So, by truncating those singular values whose “scaling” properties are heuristically deemed insignificant, we may obtain a low-rank approximation of A. Now, let us recall (3.1) and observe that the r singular values $\sigma_i = \langle W_i, A \rangle$ are essentially the coordinates of A with respect to an orthonormal basis $\alpha = (W_i)_{i=1}^{r}$. Hence, we see that the rank 1 matrices $W_i = U_i V^T$ constitute such a basis and that they span an r dimensional space. So, by performing a low-rank approximation of A, we induce a projection of A into $k < r$ dimensional space, such that its representation A_k in this lower dimensional setting is coordinatized by the first k singular values. Furthermore, this k dimensional space is spanned by $\alpha' = (W_i)_{i=1}^{k}$. So, now we have

$$A_k = U_{m\times k} \Sigma_{k\times k} V^t_{k\times n} = \sum_{i=1}^{k} U_{\bullet i} \sigma_i V^T_{\bullet i}, \quad k < r \tag{3.12}$$

where $\gamma' = (U_{\bullet i})_{i=1}^{k}$ and $\beta' = (V_{i\bullet})_{i=1}^{k}$ are bases that span the lower dimensional document and term spaces respectively. So representations of the document and term vectors in these spaces are akin to those coordinate vectors in (3.7) and (3.10), but with their last $r - k$ coordinate entries truncated. Hence the coordinate vectors $[D_i]_\gamma$ and $[T_j]_\beta$ are projected into a lower k-dimensional setting. Consequently, the effect this projection has on these vectors is that their directions are altered and constrained into fewer ones, confining them more tightly. This also has the effect of possibly binding similar semantic dimensions more closely, thereby revealing latent semantic connections between them, but whether this actually happens, depends largely on our choice of k. One way of choosing k is to measure the contribution of each singular value with the following formula

$$f_k = \frac{\sigma_k^2}{\sum_{i=1}^{r} \sigma_i^2}, \quad k = 1, \ldots, r, \tag{3.13}$$

where $f_k \in (0, 1)$ is a variation score for each singular value. There are of course other heuristics[2] for determining k, but those are beyond the scope of this article.

3.2 *Assessing Document Similarities*

One of the main goals of latent semantic indexing or LSI is to determine which documents in a given collection are similar. As we saw in the previous subsection, the low rank approximation A_k of the term document matrix A is vital for performing such a task. From this we can proceed to examine the coordinate vectors representing the document vectors in the lower dimesional document space $\mathcal{D}'$ spanned by the columns of the matrix $U_{m\times k}$. Observe that these coordinate represenations reside in the columns of the matrix $(\Sigma V^t)_{k\times n}$. So comparing documents in this space is tantamount to examining the columns of this matrix. These are column vectors, so one way of assessing their proximity to one another is by measuring the distance between them using a euclidean norm, or by measuring the angles between them. So focusing on the first approach, we

[2] Moody T Chu. On the statistical meaning of truncated singular value decomposition. *Preprint*, 2001; Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51:981–930, 2006; and S. Urien, O. Wolkenhauer, Electrical Engineering, and Electronics. *Singular Value Decomposition for Genome-wide Expression Data Processing and Modelling*. UMIST, 2002

observe that the i-th column of matrix $(\Sigma V^t)_{k\times n}$ has form:

$$\left(\Sigma V^t\right)_{\bullet i} = \left(\sigma_\bullet V^t_{\bullet i}\right) = \begin{pmatrix} \sigma_1 \left(V^t\right)_{1i} \\ \sigma_2 \left(V^t\right)_{2i} \\ \vdots \\ \sigma_k \left(V^t\right)_{ki} \end{pmatrix}, i = 1, \ldots, n. \qquad (3.14)$$

Then, applying the euclidean norm to the pair-wise difference of each column vector, we have

$$\texttt{sim}_1\left(D_i, D_j\right) = \left\|\left(\Sigma V^t\right)_{\bullet i} - \left(\Sigma V^t\right)_{\bullet j}\right\|_2 = \left\|\Sigma\left(V^t_{\bullet i} - V^t_{\bullet j}\right)\right\|_2. \qquad (3.15)$$

This has the effect of measuring the euclidean distance between each column vector (document vector) residing in $\mathcal{D}'$. Distance score values that are close to zero indicate that the documents are very similar in content, whereas higher values indicate a difference in content. Now, for the second approach, we attempt to compute the angles between these column vectors, but first taking into account the possibility that the magnitude of each column vector may vary. Thus it is appropriate to perform normalization by multiplying each column vector by the reciprocal of its euclidean norm so that the cosine angle between pairs of column vectors is

$$\texttt{sim}_2\left(D_i, D_j\right) = \frac{\left\langle (\Sigma V^t)_{\bullet i}, (\Sigma V^t)_{\bullet j}\right\rangle}{\left\|(\Sigma V^t)_{\bullet i}\right\| \left\|(\Sigma V^t)_{\bullet j}\right\|.} \qquad (3.16)$$

There is a problem with the first similarity metric $\texttt{sim}_1$ in that it doesn't take into account the length of each document being compared. As a consequence, it's possible that two documents with very similar content obtain a very a high value for their distance score. Reason being, the relative distributions of the terms may be nearly identical for both documents, but the absolute term frequencies of one document may be significantly larger[3].

4 Summary

- The term-document matrix A is an $m \times n$ matrix whose m rows represent terms and whose n columns represent documents. The entries in this matrix correspond to counts of the vocabulary terms. Sometimes weighting methods[4] are used on the terms to increase accuracy.
- The rows of A are the coordinates of the term vectors with respect to the standard basis for $\mathbb{R}^m$, and the columns of V are the coorindates of document vectors with respect to the standard basis for $\mathbb{R}^n$.
- The rank r of the matrix A is the maximal number of linearly independent columns spanning its column space (document space). Each of these r columns constitute a basis spanning the column space of A. The rank r also corresponds to the dimension of this space. Conversely, the rank r is the maximal number of linearly independent rows spanning the row space (terms space) of A.

[3] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008

[4] T. Roelleke. *Information Retrieval Models: Foundations and Relationships.* Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2013

- The SVD of A expresses A with respect to an orthonormal basis comprised of r rank one matrices. This representation of A helps us understand how many dimensions capture most of the variation in A. We study this variation by examining the r singular values, then proceed to retain only those singular values whose variation scores are deemed significant. The singular values in the SVD are in non-increasing order. Also, the SVD of A produces orthonormal bases for the four most important subspaces associated with A, amongst them are the row space (term space) and column space (document space).
- If the SVD of A is $U\Sigma V^t$, then the first r columns of U form an orthonormal basis spanning the column space (document space) of A. Similarly, the r rows of V^t form an orthonormal basis spanning the row space of A (term space).
- By performing a low rank approximation of A using knowledge obtained from its singular values, we may find a representation of the column space (row space) of A in a lower dimensional setting that better expresses the relationships the between documents (terms).
- The low rank approximation of A is lossy, but it has the advantage of potentially reducing noise that contanimates the informational content of A. This type of approximation is lossy in the sense that it induces a projection of A into a lower dimensional setting. Projections are not invertible, meaning that once they're applied to a vector (matrix), we cannot apply another transformation to its image to recover the vector's (matrix) original form (pre-image). With that in mind, determining how many singular values to retain is an art. One must be judicious with how this is done.
- After obtaining an appropriate low rank approximation of A , we may then begin to discover similarities between documents by measuring the proximity of their vector representations to one another in a lower dimensional setting. We can achieve this by using the vector norm, or obtaining the cosine angle between the document vectors residing in this lower dimensional document space. We can also measure the proximity of terms with each other using similar methods described above.

5 *An Example*

Consider the term-document tabulation for the following documents:

	D_1	D_2	D_3	D_4	D_5
`Doctor`	2	0	8	6	0
`Car`	1	6	0	1	7
`Nurse`	5	0	7	4	0
`Hospital`	7	0	8	5	0
`Wheel`	0	10	0	0	7

Table 1: Term-Document Tabulation: data borrowed from Kirk Baker's SVD tutorial.

The matrix representation of this tabulation is then given by

$$A = \begin{pmatrix} 2 & 0 & 8 & 6 & 0 \\ 1 & 6 & 0 & 1 & 7 \\ 5 & 0 & 7 & 4 & 0 \\ 7 & 0 & 8 & 5 & 0 \\ 0 & 10 & 0 & 0 & 7 \end{pmatrix}.$$

The constituents of the singular value decomposition of A are given by:

$$U = \begin{pmatrix} -0.54 & 0.06 & 0.82 & 0.11 & -0.12 \\ -0.10 & -0.59 & -0.11 & 0.79 & 0.06 \\ -0.52 & 0.06 & -0.21 & -0.12 & 0.81 \\ -0.64 & 0.07 & -0.51 & -0.06 & -0.56 \\ -0.06 & -0.80 & 0.09 & -0.59 & -0.04 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 17.92 & 0 & 0 & 0 & 0 \\ 0 & 15.17 & 0 & 0 & 0 \\ 0 & 0 & 3.56 & 0 & 0 \\ 0 & 0 & 0 & 1.98 & 0 \\ 0 & 0 & 0 & 0 & 0.35 \end{pmatrix},$$

and

$$V^t = \begin{pmatrix} -0.46 & -0.07 & -0.74 & -0.48 & -0.06 \\ 0.02 & -0.76 & 0.10 & 0.03 & -0.64 \\ -0.87 & 0.06 & 0.28 & 0.4 & -0.04 \\ -0.00 & -0.60 & -0.22 & 0.33 & 0.69 \\ -0.17 & -0.23 & 0.57 & -0.70 & 0.32 \end{pmatrix}.$$

Examining Σ we are able to see that semantic dimensions 1 and 2 are the most important due to the relatively large magnitudes of their associated singular values, $\sigma_1 = 17.92$ and $\sigma_2 = 15.7$, respectively. Using (3.13), we plot the variation contribution from each singular in figure 5.1. We determine that $f_1 \approx 0.57$ and $f_2 \approx 0.41$, so the first two singular values account for about 98% of the variation in A. That being said, we retain the first two singular values and discard the rest. So, now the rank 2 approximation of A is given by

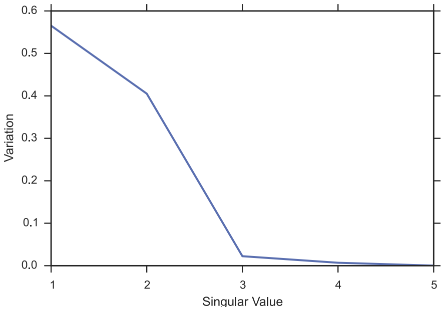

Figure 5.1: The contribution of each singular value is measured according to (3.13)

$$A_2 = U_{5\times2}\Sigma_{2\times2}V^t_{2\times5}.$$

The constituents of the triadic decomposition above are

$$U_{5\times2} = \begin{pmatrix} -0.54 & 0.06 \\ -0.10 & -0.59 \\ -0.52 & 0.06 \\ -0.64 & 0.07 \\ -0.06 & -0.80 \end{pmatrix}, \; V^t_{2\times5} = \begin{pmatrix} -0.46 & -0.07 & -0.74 & -0.48 & -0.06 \\ 0.02 & -0.76 & 0.10 & 0.03 & -0.64 \end{pmatrix}, \; \Sigma_{2\times2} = \begin{pmatrix} 17.92 & 0 \\ 0 & 15.17 \end{pmatrix}.$$

It is evident that the document vectors associated with documents 1-5 are projected into a 2 dimensional space spanned by the columns of $U_{5\times2}$, and that the vector representations of the documents in this space are contained in the columns of $\Sigma V^t_{2\times5}$. By inspecting the columns of $V^t_{2\times5}$, we surmise that documents 1 and 4 are similar, and the same can be said for documents 2 and 5.

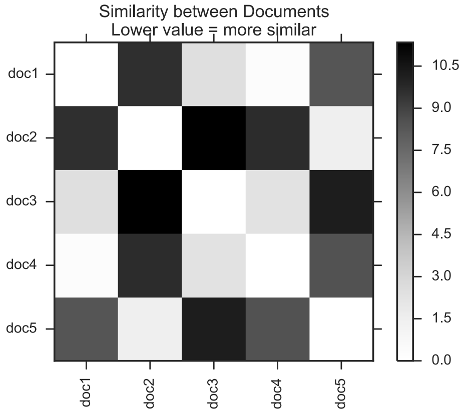

Figure 5.2: Heatmap displaying similarities between documents. The metric $\mathtt{sim}_1$ was used.

We employ similarity metric 1 to calculate the distance between the column vectors of $\Sigma V^t_{2\times 5}$ and colorize their scores in a heat map above. The similarity scores are actually tabulated in Table 2.

	D_1	D_2	D_3	D_4	D_5
D_1	0	9.59	2.42	0.17	8.34
D_2	9.59	0	11.34	9.69	1.37
D_3	2.42	11.34	0	2.25	10.2
D_4	0.17	9.69	2.25	0	8.45
D_5	8.34	1.37	10.2	8.45	0

Table 2: similarity scores for document vectors.

Meantime, the document vectors in the reduced document space are plotted below.

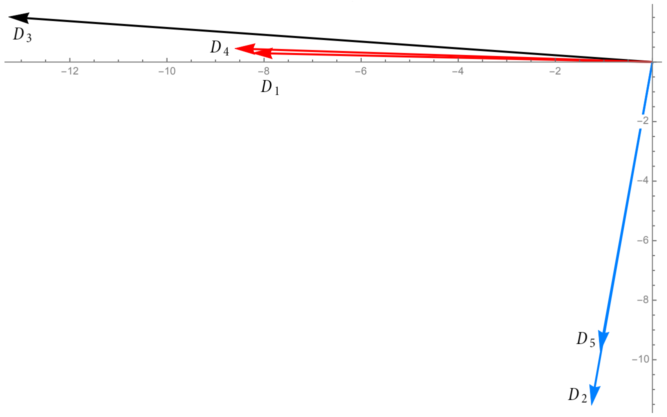

Figure 5.3: Plot of document vectors in space spanned by the first two columns of U.

As an exercise, we encourage the reader to compute the cosine similarity scores for documents, and to develop proximity metrics for word co-occurences.

6 Appendix: Linear Algebra

6.1 Definitions

Definition 6.1. Let $A \in \mathcal{M}_{m \times n}(\mathbb{R})$, that is, an $m \times n$ matrix with entries from the scalar field $\mathbb{R}$. Then the ***transpose*** of A, denoted by A^t is obtained by interchanging the rows of A with its columns. Hence the entries of A^t, satisfy $(A^t)_{ij} = A_{ji}$.

Definition 6.2. Let $A \in \mathcal{M}_{m \times n}(\mathbb{C})$, then the ***Frobenius Norm***, denoted by $\|A\|_F$ is defined by the equations:

$$\|A\|_F^2 = \sum_{i,j} \left|A_{ij}\right|^2 = \sum_i \|A_{i\bullet}\|_2^2 = \sum_j \left\|A_{\bullet j}\right\|_2^2 = \text{trace}\left(A^\star A\right).$$

The dot in $A_{i\bullet}$ indicates an aggregation over the index j, while the dot in $A_{\bullet j}$ indicates an agregration over the index i. Also $A^\star$ is the conjugate transpose of A. If A takes only real entries, then $A^\star = A^t$.

Definition 6.3. Let $\beta = \{v_1, v_2, \ldots, v_n\}$ be an ordered orthonormal basis for a vector space $\mathcal{V}$, then every vector in $x \in \mathcal{V}$ may be uniquely expressed as

$$x = \sum_{i=1}^{n} \langle v_i, x \rangle \, v_i .$$

This type of linear combination is called the ***Fourier expansion*** of x with respect to an orthonormal basis β. The scalars $\langle v_i, x \rangle$ formed by the inner product $\langle \cdot, \cdot \rangle$ are called ***Fourier coefficients***. Consequently, we define the ***coordinate representation (or coordinate vector)*** of x with respect to the orthonormal basis β to be

$$[x]_\beta = \begin{pmatrix} \langle v_1, x \rangle \\ \langle v_2, x \rangle \\ \vdots \\ \langle v_n, x \rangle \end{pmatrix}$$

where the entries in this column vector are the ***coordinates of** x **with respect to a basis** β*.

Example 6.4. Suppose we have a term document matrix given by

$$A = \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 2 \end{pmatrix}$$

Where the columns of A represent documents D_1, D_2 (in that order). Also, the rows of A represent semantic units or terms T_1, T_2, T_3 (in that order). Then, the matrix A has an order $r = 2$ SVD given by

$$U = \begin{pmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sqrt{6} & 0 \\ 0 & 2 \end{pmatrix}, \quad V^t = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

Now, let us compute ΣV^t and verify that its columns are the coordinates of the document vectors D_i with respect to a basis comprised of the columns in U. Indeed, observe that

$$\Sigma V^t = \begin{pmatrix} \sqrt{3} & \sqrt{3} \\ -\sqrt{2} & \sqrt{2} \end{pmatrix}.$$

Then, since the columns of U constitute an orthonormal basis $\gamma = (U_{\bullet 1}, U_{\bullet 2})$, we may represent the document vectors D_1 and D_2, each as unique linear combination of the vectors in γ, moreover such linear combinations are Fourier expansions, thereby making it easy to resolve their coordinates in the space spanned by γ. Thus,

$$D_i = \sum_{k=1}^{2} \langle D_i, U_{\bullet k} \rangle \, U_{\bullet k}$$

The bullet in $V_{1\bullet}$ serves to indicate an aggregation over the column index, and so $V_{1\bullet}$ is representative of the first row in V. Similarly, $V_{\bullet 1}$ is representative of the first column in V, where the position of the bullet indicates an aggregation of the row index. This notation merely serves to help us distinguish row and column vectors.

and so the coordinate vectors of documents D_i with respect to the basis γ are essentially:

$$[D_i]_\gamma = \begin{pmatrix} \langle D_i, U_{\bullet 1} \rangle \\ \langle D_i, U_{\bullet 2} \rangle \end{pmatrix}$$

where the entries in this column vector correspond to the entries in the i-th column of ΣV^t. To see this more concretely, we compute the coordinates of document vector D_1. Indeed,

$$\langle D_1, U_1 \rangle = \left\langle \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix} \right\rangle = \frac{2}{\sqrt{3}} + \frac{1}{\sqrt{3}} + 0 = \sqrt{3}$$

and

$$\langle D_1, U_2 \rangle = \left\langle \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\rangle = -\sqrt{2}.$$

So,

$$[D_1]_\gamma = \begin{pmatrix} \sqrt{3} \\ -\sqrt{2} \end{pmatrix}, \quad [D_2]_\gamma = \begin{pmatrix} \sqrt{3} \\ \sqrt{2} \end{pmatrix}.$$

Thus

$$\Sigma V^t = \begin{pmatrix} [D_1]_\gamma & [D_2]_\gamma \end{pmatrix}.$$

Now, let us compute $U\Sigma$ and verify that its rows are the coordinates of the term vectors T_j with respect to a basis β comprised of the columns in V (or rows of V^t). Indeed, observe that

$$U\Sigma = \begin{pmatrix} \sqrt{2} & -\sqrt{2} \\ \sqrt{2} & 0 \\ \sqrt{2} & \sqrt{2} \end{pmatrix}.$$

Now, since the rows of V constitute an orthonormal basis $\beta = (V_{1\bullet}, V_{2\bullet})$, we may represent the term vectors T_1, T_2, T_3 as a fourier expansions over these basis vectors, hence

$$T_j = \sum_{k=1}^{2} \langle T_j, V_{k\bullet} \rangle \, V_{k\bullet},$$

again making it easy to resolve their coordinates in the space spanned by β. Thus, the coordinate vector of T_j with respect to β is given by

$$\left[T_j\right]_\beta = \begin{pmatrix} \langle T_j, V_{1\bullet} \rangle \\ \langle T_j, V_{2\bullet} \rangle \end{pmatrix}.$$

Hence, after some calculation, we obtain

$$[T_1]_\beta = \begin{pmatrix} \sqrt{2} \\ -\sqrt{2} \end{pmatrix}, \quad [T_2]_\beta = \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix}, \quad [T_3]_\beta = \begin{pmatrix} \sqrt{2} \\ \sqrt{2} \end{pmatrix}.$$

From this we see that the entries in each column vector $\left[T_j\right]_\beta$ correspond to entries in the j-th row of $U\Sigma$, thereby demonstrating that the rows of $U\Sigma$ are indeed the coordinates of the term vectors in the space spanned by the rows of V.

References

[1] Kirk Baker. Singular value decomposition tutorial, January 2013.

[2] Moody T Chu. On the statistical meaning of truncated singular value decomposition. *Preprint*, 2001.

[3] S.H. Friedberg, A.J. Insel, and L.E. Spence. *Linear Algebra*. Featured Titles for Linear Algebra (Advanced) Series. Pearson Education, 2003.

[4] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Matrix Analysis. Cambridge University Press, 2012.

[5] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[6] E.V. Munson, C. Nicholas, and D. Wood. *Principles of Digital Document Processing: 4th International Workshop, PODDP'98 Saint Malo, France, March 29–30, 1998 Proceedings*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2003.

[7] T. Roelleke. *Information Retrieval Models: Foundations and Relationships*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2013.

[8] G. Strang. *Linear Algebra and Its Applications*. Thomson, Brooks/Cole, 2006.

[9] S. Urien, O. Wolkenhauer, Electrical Engineering, and Electronics. *Singular Value Decomposition for Genome-wide Expression Data Processing and Modelling*. UMIST, 2002.

[10] Vladislav D Veksler, Ryan Z Govostes, and Wayne D Gray. Defining the dimensions of the human semantic space. In *30th Annual Meeting of the Cognitive Science Society*, pages 1282–1287, 2008.

[11] Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51:981–930, 2006.