Representing and Accessing [Textual] Digital Information (COMS/INFO 630), Spring 2006
Lecture 17: The Singular Value Decomposition
4/4/06
Lecturer: Lillian Lee
Scribes: Nicolas Hamatake (nh39), Min Park (mp257)

1 Introduction

1.1 Overview

In today's lecture we will finally state the *Singular Value Decompotion* 'theorem'. To build some intuition for it, we will continue exploring the underlying geometric interpretations of *Matrix Theoretic Corpus Characterizations*. We wish to develop a general way of succinctly describing properties inherent in some corpora using matrices. Since we already know of a vector space representation for documents, we would be interested in characterizing the *spread* of the document vectors in a geometric sense. We will revisit two possibilities for characterizations that were discussed last time, and consider two more today[1]:

1. Rank(**D**);
2. The convex hull of the column vectors of **D**
3. The convex hull of the column vectors of **D** and - **D**
4. Mapping a *unit* (*n-1*)*-sphere* ($\subset \mathbb{R}^n$) by some **D**, as a linear operator

Having spent nearly two lectures in algebra and geometry with only sketchy references to its application to information retrieval, the lecture concludes by initiating a discussion of the implications of *SVD* from an IR perspective.

1.2 Inherent Properties of the Corpus—No Queries Involved

The conclusion of the lectures related to relevance feedback brought forth the question, *can corpora be characterized prior to having feedback information?* Until the beginning of last lecture, most of our references to corpora have been with respect to some sort of relevance information related to some query (including the query itself). An exception would be the *idf*.

Should we expect that there are properties "inherent" of a corpus? Is it reasonable to think that corpora have structure? It certainly seems plausible that structures of corpora would be formed and maintained naturally and/or intentionally. For instance, the creator(s) of a corpus might intend to perform some process (such as information retrieval) on it, or expect that others would.

To study structural properties of corpora, we began with a search for a good representation of the corpus—a succinct *and* informative summary of it. Last time, new and also some familiar vector space notation was discussed, along with two potential corpus characterizations.

[1] See Section 2.1 for explanation of the notation presented.

2 Matrix-Theoretic Corpus Characterization

2.1 Review of Basic Notation

Assume there are *n* documents in the corpus with vector representations, $\vec{d}^{(1)},\ldots,\vec{d}^{(n)}$;
Assume an *m*-term vocabulary;
Then, the *feature term-document matrix* is an *m* by *n* matrix denoted by

$$\mathbf{D}=\begin{bmatrix} | & & | \\ \vec{d}^{(1)} & \ldots & \vec{d}^{(n)} \\ | & & | \end{bmatrix}.$$

Superscripts will be used to denote matrix columns.
Subscripts will be used to denote vector components.

2.2 Characterizations

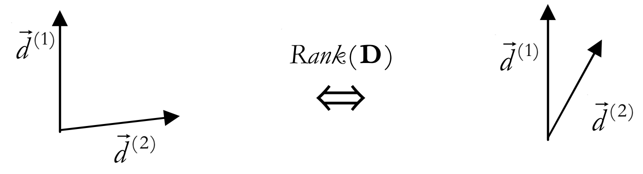
The goal is to succinctly characterize the corpus in terms of the variability of its documents. With respect to of our matrix framework and its notation, we are interested in the 'spread' of the $\vec{d}^{(i)}$'s. Two descriptions of the corpus characterizations introduced in the previous lecture are reviewed, and two more are discussed. Each new consideration is intended to refine the previous corpus description such that it might be a more informative characterization. A guiding example is introduced in Section *2.2.1.2* which will be referenced throughout this section.

2.2.1 Two Characterizations from Last Lecture:

1. The dimension of all possible linear combinations of the $\vec{d}^{(i)}$'s

Rank(**D**) is a number which is equal to the dimension of the span (all possible linear combinations) of $\left\{\vec{d}^{(i)}\right\}_{i=1}^{n}$; it can thought of as a measure of complexity in the corpus structure. Mathematically, the rank of a matrix describes the number of its linearly independent columns, and in the non-full-rank case implies that some are linear combinations of the others. In the context of document vectors, the rank of **D** would reveal how many of its columns are sufficient to describe all *n* columns representing the documents in the corpus. So if *Rank*(**D**) is 'relatively' close to *n*, one might infer that the corpus is complex. So perhaps we can consider the rank as some measure of variability among the documents, particularly since we are interested in developing some geometric intuition.

Since this corpus characterization is just one real number, it is very succinct. However, desired information is lost. For instance, if it is the case that the feature-document space represented by **D** is spanned by just two collinear document vectors $\vec{d}^{(1)}$ and $\vec{d}^{(2)}$, *Rank*(**D**)=2 regardless of the configuration of the two vectors and so will not differentiate between the following two corpora:



On the left hand side of the above diagram, the represented corpus is more spread out than the one represented on the right hand side. In this sense, they are different and Rank(**D**) is not 'fine grained' enough to capture this point.

Could there be a different characterization that is still succinct, and still retains more information about the corpus structure? Since *Rank*(**D**) in indirectly considering a set of *all possible* linear combinations of a basis set of vectors for the feature document space by describing its dimension, maybe a subset of the set of linear combinations (henceforth denoted by $\mathcal{a}$) could be more informative.

2. One Particular Set of Linear Combinations: $\mathcal{a}_0$

To consider the new idea of examining properties of a portion of span(**D**), we can consider coefficient vectors denoted by

$$\vec{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \in \mathbb{R}^n ,$$

that are associated with a restricted set of linear combinations of the $\vec{d}^{(i)}$'s. Furthermore, we take an intuitively unusual perspective of **D** as an operator on these coefficient vectors[2].

.

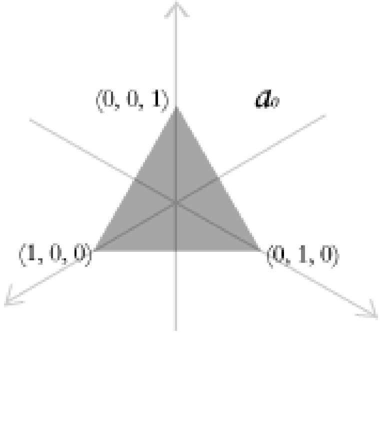
Now we consider a special portion of $\mathbb{R}^n$ that corresponds to *fractional assignment* of the columns of the feature term-document matrix. That is,

$$\mathcal{a}_0 = \left\{ \vec{\alpha} \in \mathbb{R}^n : \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i = 1 \right\},$$

To guide our discussion, consider the case in which n = 3**,** and m = 2. Then,

$$\mathcal{a}_0 = \left\{ \vec{\alpha} \in \mathbb{R}^3 : \alpha_i \geq 0, \sum_{i=1}^{3} \alpha_i = 1 \right\},$$

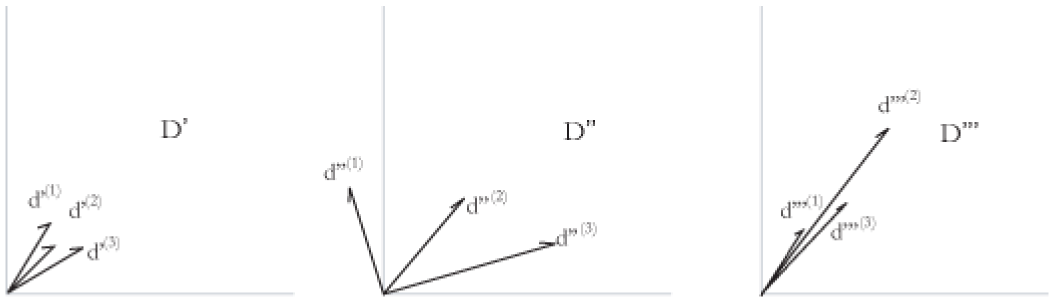
and can be geometrically illustrated in the following diagram.



Note: $\mathcal{a}_0 \subset \mathcal{a}$.

[2] It is usually more 'natural' to think of the coefficient vectors as the linear operator on data (**D**) instead.

Also consider a visual illustration in $\mathbb{R}^2$ of three examples of feature-document matrices, $\mathbf{D}', \mathbf{D}'', \mathbf{D}''',\ \subset\ \mathbb{R}^{2\times3}$ (each **D** representing a set of three two-term documents):

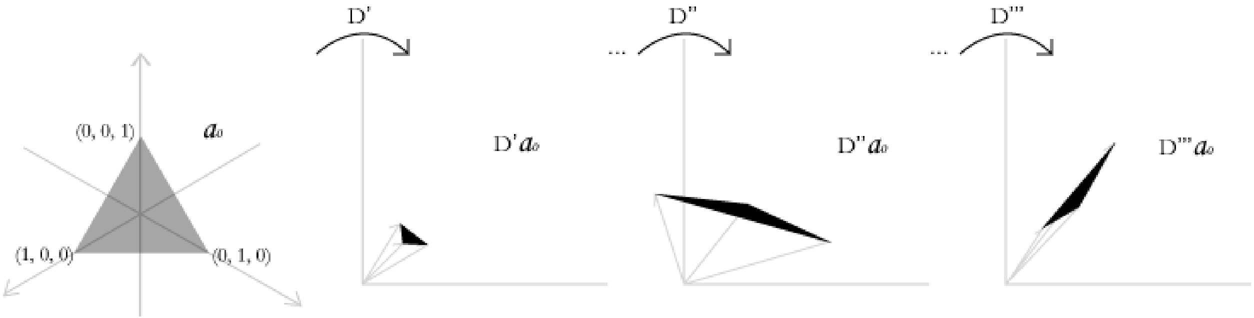


- $\mathbf{D}'$ is an example of a set of documents with little spread;
- $\mathbf{D}''$ is an example of a set of documents of more spread;
- $\mathbf{D}'''$ is an example of a set of documents also with little spread, but high variation along a single direction.

Then our hope should be that an appropriate characterization will reveal that

- $\mathbf{D}'$ and $\mathbf{D}''$ different[3]
- $\mathbf{D}''$and $\mathbf{D}'''$ are similar[4]
- $\mathbf{D}''$ and $\mathbf{D}'''$ are different.

If we apply these **D**'s to a_0, we will get $\mathbf{D}'a_0$, $\mathbf{D}''a_0$ and $\mathbf{D}'''a_0$, the respective convex hulls of $\mathbf{D}', \mathbf{D}'',$ and $\mathbf{D}'''$, (represented by black triangles in the illustration below):



$\mathbf{D}': \mathbb{R}^3 \to \mathbb{R}^2$
$a_0 \mapsto \mathbf{D}'a_0$

$\mathbf{D}'': \mathbb{R}^3 \to \mathbb{R}^2$
$a_0 \mapsto \mathbf{D}''a_0$

$\mathbf{D}''': \mathbb{R}^3 \to \mathbb{R}^2$
$a_0 \mapsto \mathbf{D}'''a_0$

From last time, our first-glance comparison of $\mathbf{D}'a_0$ and $\mathbf{D}''a_0$ suggested that we should refine our corpus characterization with a notion of area with respect to these convex hulls. We observed from these two mappings of a_0 that we could differentiate structures of the same rank. This description successfully characterizes $\mathbf{D}'$ and $\mathbf{D}''$as different; feature vectors of $\mathbf{D}'$ have little

[3] 'Different' under our notion of spread

[4] 'Similar' under our notion of spread

spread and also a small convex hull with respect to a_0, while those of $\mathbf{D}''$ have greater spread, with a larger convex hull. We wondered if this situation would generalize.

Unfortunately, this possibility was abandoned by observing the counterexample provided by the convex hull of $\mathbf{D}'''$. While the feature vectors of $\mathbf{D}'''$ have little spread, they have a large convex hull, similar in size to that of $\mathbf{D}''$. In other words, this particular characterization with a_0 would treat $\mathbf{D}''$ and $\mathbf{D}'''$ as similar, and $\mathbf{D}'$ and $\mathbf{D}'''$ as different, which does not coincide with our intended notion of "different".

Could we refine our characterization even more? The observations from our low-dimensional example suggest that we should, and hints that it could be of benefit to have one that captures some notion of *direction*.

2.2.2 *Two more characterizations*

3. A Particular Set of Linear Combinations: a_1

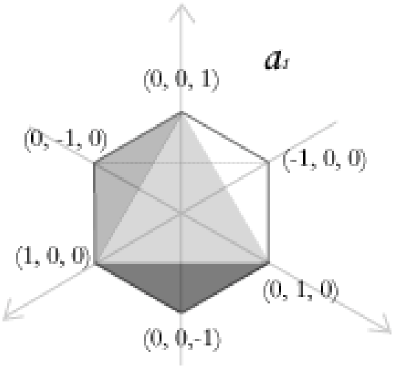
What if we consider a slightly larger portion of $\mathbb{R}^n$ than a_0? For instance,

$$a_1 = \left\{ \vec{\alpha} \in \mathbb{R}^n : \alpha_i \geq 0, \sum_{i=1}^{n} |\alpha_i| = 1 \right\}$$
$$= \left\{ \vec{\alpha} \in \mathbb{R}^n : \|\vec{\alpha}\|_1 = 1 \right\}$$

This set represents an *(n-1)*-octahedron in $\mathbb{R}^n$. In terms of our concrete example introduced in the previous section, we have

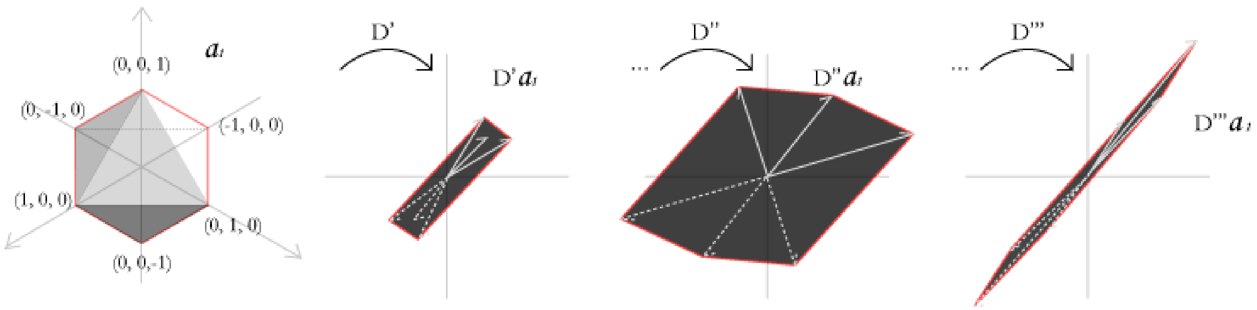
$$a_1 = \left\{ \vec{\alpha} \in \mathbb{R}^3 : \|\vec{\alpha}\|_1 = 1 \right\},$$

which can be geometrically illustrated in the following diagram:



Note: $a_0 \subset a_1 \subset a$.

To map a_1 by $\mathbf{D}', \mathbf{D}'', \mathbf{D}'''$, we can first map the six vertices of the *2*-octahedron as marked above. The mapping is the convex hull of the column vectors of **D** and **–D;** hence, **D'** a_1, **D"** a_1, **D'''** a_1 should be hexagons in $\mathbb{R}^2$ (represented by the semi-transparent areas in the illustration below):



$\mathbf{D}': \mathbb{R}^3 \to \mathbb{R}^2$ $\quad$ $\mathbf{D}'': \mathbb{R}^3 \to \mathbb{R}^2$ $\quad$ $\mathbf{D}''': \mathbb{R}^3 \to \mathbb{R}^2$

$a_1 \mapsto \mathbf{D}'a_1$ $\quad$ $a_1 \mapsto \mathbf{D}''a_1$ $\quad$ $a_1 \mapsto \mathbf{D}'''a_1$

In our example, we can see that we observe the desired effect. $\mathbf{D}'a_1$ and $\mathbf{D}'''a_1$ are both 'skinny' in comparison to $\mathbf{D}''a_1$ which is 'fat'. This is in agreement with the relative spreads among $\mathbf{D}', \mathbf{D}''$, and $\mathbf{D}'''$. We have successfully characterized these **D's** with a notion of *directed area.*

Still, we should note a couple of problems with this characterization. To highlight a less serious problem first: we observe that **D'** is a degenerate linear operator. It maps an object with *six* vertices to an object with only *four* vertices. The mapping of (0, 1, 0) and (0, -1, 0) ∈ a_1 by $d'^{(2)}$ is contained within the supposed a hexagon (actually a rectangle), $\mathbf{D}'a_1$. Since the degenerate mapping still captures the consensus direction of the corpus, this is not a severe problem for the purpose of corpus characterization, although we might only be able to recover partial information about **D'** if we were interested.

A more serious issue with this characterization is that it is not succinct. In the above example, we have three documents vectors in $\mathbb{R}^2$ for each **D**, but require six vectors in $\mathbb{R}^3$ to obtain its $\mathbf{D}a_1$ (hexagonal) characterization. Descriptions of polyhedra are often complex and we could imagine that this characterization could be more complicated than **D** itself.

We have gained the desired refinement for our characterization, but in turn, lost the succinctness that we had before. Could there be a summary that is as characterizing as, but more concise than polyhedral mappings by the feature-term document matrix **D**?

4. Another Particular Set of Linear Combinations: a_2

We have considered the dimension of all possible linear combinations, fractional assignments, and their 'negatives' (1-norm = 1); next, let us consider the following:

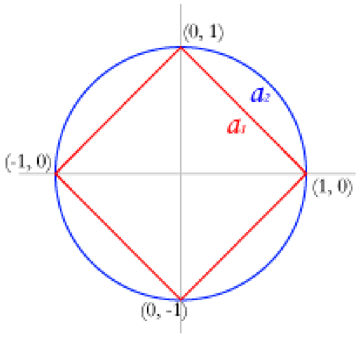
$$\begin{aligned} a_2 &= \left\{ \vec{\alpha} \in \mathbb{R}^n : \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i^2 = 1 \right\} \\ &= \left\{ \vec{\alpha} \in \mathbb{R}^n : \|\vec{\alpha}\|_2 = 1 \right\} \end{aligned}$$

Note: (1) a_2 represents a unit (*n-1*)-sphere in $\mathbb{R}^n$;

(2) $a_0, a_1 \not\subset a_2$;

(3) a_1's vertices are contained in a_2; $a_1 \cap a_2 = \left\{ |e_i| \in \mathbb{R}^n, \forall i \right\}$.

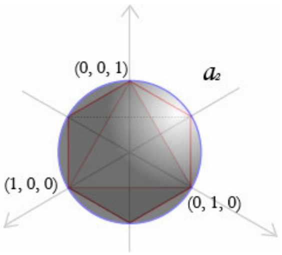
For example, a_1 and a_2 in $\mathbb{R}^2$:



We could consider a_2 to have a '**smoothing effect**' for a_1. To return to our example:

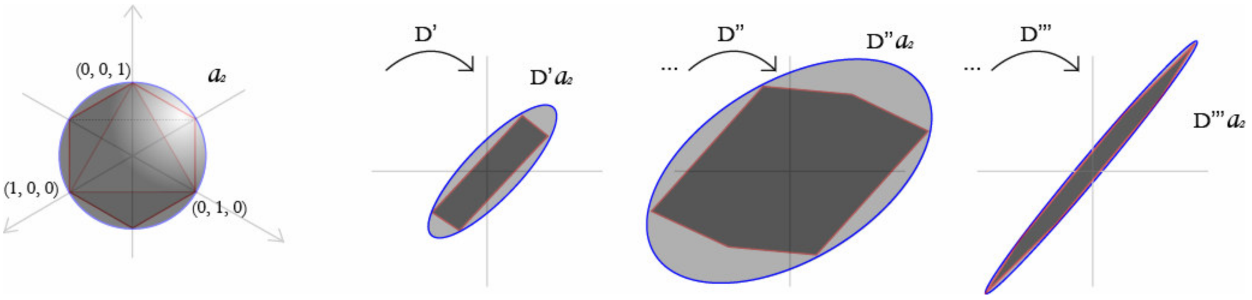
$$a_2 = \left\{ \vec{\alpha} \in \mathbb{R}^3 : \|\vec{\alpha}\|_2 = 1 \right\},$$

which is geometrically illustrated in the following diagram.

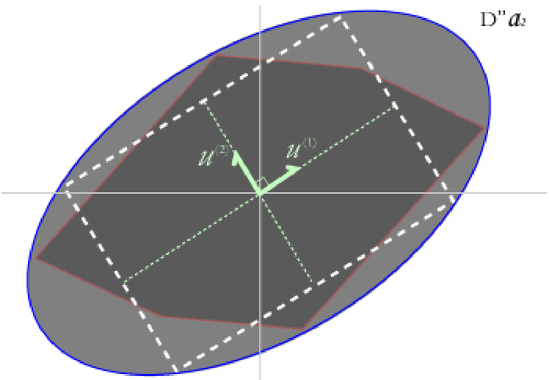


This 'smoothing effect' means that the mapping of a_1 will be smoothed into a hyperellipsoid. An added benefit is that this characterization would not lead to degenerate mappings; linear mappings of a sphere are limited to ellipsoids in the target space.

We can see the smoothing effect again in following illustrations of the mappings of a_2 by **D'**, **D''**, and **D'''** (superimposed on mappings of a_1):



Moreover, ellipsoids have much more succinct descriptions than polyhedra. For instance, consider **D''** a_2 (superimposed on the mapping of a_1):



This ellipsoid can be described by the two directions of its principal axes, given by a pair of orthonormal vectors $u^{(1)}$ and $u^{(2)}$, ordered by their lengths, $\sigma_1 > \sigma_2$. In comparison to describing the six points of a hexagon, this is certainly more concise. An r-dimensional hyperellipsoid can be described by:

1. A set of r orthonormal vectors that describe its axes, $\vec{u}^{(1)},...,\vec{u}^{(r)}$;
2. A respective ordered set of scalars $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$ to describe the lengths of the $\vec{u}^{(l)}\text{'}s$.

To summarize, we now have a relatively terse[5] characterization of our feature term-document matrix **D** that also maintains the desired notion of directed area as in the previous 'a_1 characterization'.

There is a minor technical problem if there are non-distinct σ_l's; their respective $\vec{u}^{(l)}\text{'}s$ would not be unique for that corpus. In our example, imagine that the sphere is mapped to a disc. This only a minor and technical issue since dimensions we are not interested in unique identification of a corpus—only a characterization of its structure. In some sense, if there *are* some non-distinct σ_l's, that itself is a property of the associated corpus structure.

2.3 The '*Magic*' of Singular Value Decomposition
—In Context of Our Discussion

Which coefficient vectors in a_2 are mapped to the axes of the hyperellipsoid $\mathbf{D}a_2$? Could they be meaningful to us since their images happen to characterize our feature term-document matrix? How can we recover them, given **D** and a_2?

Recall from last lecture the peculiar duality between the document associations and term associations in the corpus described by Rank(**D**)=Rank($\mathbf{D}^T$). We took a rough interpretation of the *m* row vectors of **D** to describe some sort of underlying document-distribution for each of the *m* terms. We could find an orthonormal basis in $\mathbb{R}^n$ of the columns of $\mathbf{D}^T$ — the term associations in the corpus. Moreover they would be contained in a_2.

That is to say, if we any had *r* orthonormal vectors in a_2 as the columns of some matrix, its rows would be $\vec{d}^{(1)},...,\vec{d}^{(n)}$ with respect to some basis set of vectors for **D**.

[5] We hope Rank(**D**)= $r < n$, otherwise our characterization is as succinct as **D** itself.

$$\mathbf{V} = \underset{n \times r}{\begin{bmatrix} | & & | \\ \vec{v}^{(1)} & \cdots & \vec{v}^{(r)} \\ | & & | \end{bmatrix}}$$

Now consider the pre-images of the axes of $\mathbf{D}\,a_2$, $\vec{v}^{(l)} \in \mathbb{R}^n$. Since it is a linear image of a unit sphere, for all $\vec{u}^{(l)}$, there must exist a unit vector $\vec{v}^{(l)} \in \mathbb{R}^n$ such that $\mathbf{D}\,\vec{v}^{(l)} = \sigma_l \vec{u}^{(l)}$. Since the hyperellipsoid itself will have *r* axes, there are *r* pre-images to consider; how do we recover them? We could recover them by using algorithms computing the *Singular Value Decomposition*—which factors any *m* by *n* matrix such as **D** to produce three matrices of the following form:

$$\mathbf{D} = \underset{m \times r}{\begin{bmatrix} | & & | \\ \vec{u}^{(1)} & \cdots & \vec{u}^{(r)} \\ | & & | \end{bmatrix}} \underset{r \times r}{\begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{bmatrix}} \underset{r \times n}{\begin{bmatrix} - & \vec{v}^{(1)} & - \\ & \vdots & \\ - & \vec{v}^{(r)} & - \end{bmatrix}} = \mathbf{U}\Sigma\mathbf{V}^T$$

U is *the left singular matrix*, with orthonormal columns in $\mathbb{R}^m$; Σ, is a diagonal matrix of *r* singular values of **D**, and $\mathbf{V}^T$ is *the right singular matrix*, with orthonormal rows in $a_2 \subset \mathbb{R}^n$.

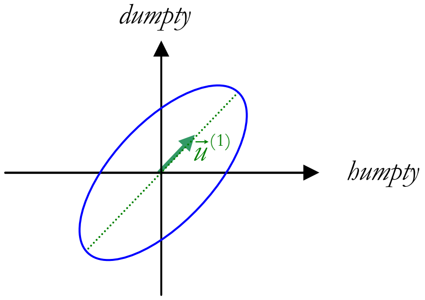
From the SVD of **D**, we have immediately recovered the pre-images of the hyperellipsoid's axes! Also, respectively $\mathbf{U}$, and Σ are the orthonomal directions for the axes of our corpus characterizing hyperellipsoid, and their lengths. Even more intriguing is that the columns $\mathbf{V}^T$ describe $\vec{d}^{(1)},\ldots,\vec{d}^{(n)}$ with respect to the basis described by the columns of $\mathbf{U}$, for reasons discussed earlier in this section

3 SVD in IR

3.1 Are the $\vec{u}^{(l)}$'s co-occurrence patterns?

What is the meaning of the SVD of **D** in context of information retrieval? Could we interpret the left singular vectors as co-occurrence patterns (since they form a basis for the subspace occupied by the document vectors)? For example, the term "humpty" and "dumpty" are likely to often occur together; then the existence of co-occurrence patterns among terms in general certainly seems reasonable.

Illustrated in context of our main example from Section 2, we might conclude that most documents (or at least one) contain "humpty" and "dumpty" with roughly the same (large) frequency.



It would be nice if could interpret $\vec{u}^{(l)}$'s as co-occurrence patterns among terms

It is tempting—especially from an IR perspective—to hope that the $\vec{u}^{(l)}$**'**s might even represent topics. These wishful interpretations would make it easy to place the *SVD* in the context of *IR* models. However, this is probably not the case since the $\vec{u}^{(l)}$'s are mutually orthogonal. Orthogonality is improbable with topics. It seems only slightly more likely for co-occurrence patterns.

Given our motivation for corpus characterization, a more credible interpretation of the set of $\vec{u}^{(l)}$'s is that it reveals directions of 'most' variation in the corpus (away from zero, not between vectors[6]). The implications of this observation to information retrieval are not immediately obvious. On the other hand, recall that the ***U****-basis vectors from the SVD* of **D** are particular because they are *ranked* by their respective singular values, which are the lengths of the axes of the hyperellipsoid.

With this remark, we will consider a mathematical theorem that restores our hope that the $\vec{u}^{(l)}$'s do give co-occurrence patterns—the important ones at least.

3.2 Eckart-Young Theorem

Consider a space spanned by the top *k* left singular vectors. Suppose we project **D** into this space. The Eckart-Young Theorem states that this will give the best rank-*k* approximation to **D,** in a precise mathematical sense.

In an information retrieval *sense*, how would approximating **D** get us closer to some "truer" representation of documents? One idea is that if two terms occur together frequently—such as "humpty" and "dumpty"—we could 'save' on a dimension by approximation, which will then identify low ranked $\vec{u}^{(l)}$'s as noise.

Then the Eckart-Young theorem allows for a less unreasonable hope that for IR, $\vec{u}^{(l)}$'s *do* give *important* co-occurrence patterns. When will this interpretation be useful, and when would it degenerate? LSI

[6] One could wonder if the 0 vector has special meaning in the data.

is an application of this theorem, which has shown cases of very good performance. In spite of this, there is still no completely satisfying basis for understanding why it does work so well—when it does.

"*LSI*"—the Eckart Young Theorem applied to IR—will be discussed next class…

4 Questions

1. Consider the following matrices:

$$A=\begin{bmatrix}1.1 & 1.2 & 3.0 & 2.1 & 0.0\\ 1.0 & 1.3 & -3.1 & -2.0 & 0.0\\ 0.1 & 0.1 & -0.1 & 0.0 & 0.0\end{bmatrix}$$

$$B=\begin{bmatrix}1.1 & 1.2 & 3.0 & 2.1 & 0.0\\ 1.0 & 1.3 & -3.1 & -2.0 & 15\\ 0.1 & 0.1 & -0.1 & 0.0 & 15\end{bmatrix}$$

a) How do you expect the SVD's of A and B to compare (no exact SVD calculations needed)?

b) How does document length affect the left-singular vectors, assuming the document vectors are not length-normalized?

c) In the general setting, suppose that we use the SVD to obtain a lower-rank approximation for the term-document matrix, and score documents using this lower-dimensional subspace. It appears that our model favors co-occurrences of terms in long documents over those in short documents. Is this good or bad, and in what ways?

2. Suppose that, in order to decrease the effect of document length on the SVD, we initially scale each document vector so that its 2-norm is one. What kind of effect would this have on scoring?

5 Answers

1a) Upon close examination, it appears that A is very close to being a rank-two matrix. Thus, one would expect A to have two left-singular vectors with corresponding singular values that are significantly larger than the others. Their directions would be approximately $\begin{bmatrix}1 & 1 & 0\end{bmatrix}$ and $\begin{bmatrix}1 & -1 & 0\end{bmatrix}$. One would expect B to have the same singular vectors, plus an additional one in the direction of $\begin{bmatrix}0 & 1 & 1\end{bmatrix}$. This is verified by examining the precise SVD's:

$$A = \begin{bmatrix} -0.703 & 0.711 & 0.033 & 0.000 & 0.000 \\ 0.711 & 0.701 & 0.055 & 0.000 & 0.000 \\ 0.016 & 0.062 & -0.998 & 0.000 & 0.000 \end{bmatrix} \times \begin{bmatrix} 5.199 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 2.315 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.052 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{bmatrix} \times \ldots$$

$$\begin{bmatrix} -0.012 & 0.643 & -0.178 & 0.745 & 0.000 \\ 0.016 & 0.765 & 0.204 & -0.611 & 0.000 \\ -0.830 & -0.021 & 0.541 & 0.134 & 0.000 \\ -0.557 & 0.039 & -0.796 & -0.232 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix}$$

$$B = \begin{bmatrix} 0.018 & 0.864 & 0.504 & 0.000 & 0.000 \\ -0.720 & -0.339 & 0.606 & 0.000 & 0.000 \\ -0.694 & 0.374 & -0.616 & 0.000 & 0.000 \end{bmatrix} \times \begin{bmatrix} 21.424 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 4.504 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.765 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{bmatrix} \times \ldots$$

$$\begin{bmatrix} -0.036 & 0.144 & 0.622 & -0.768 & 0.000 \\ -0.046 & 0.141 & 0.754 & 0.639 & 0.034 \\ 0.110 & 0.800 & -0.172 & 0.005 & 0.564 \\ 0.069 & 0.553 & -0.087 & 0.030 & -0.825 \\ -0.990 & 0.116 & -0.083 & 0.001 & 0.004 \end{bmatrix}$$

1b) Long documents have a greater influence in shaping the hyper-ellipsoid discussed in class. Thus, the left-singular vectors seem to favor the directions of the longest documents.

1c) This depends, as with the VSM model, on why long documents in the corpus are long. If they are long because they incorporate a large amount of relevant information from other documents in the corpus, then it would be good to have highly-ranked singular vectors that are similar in direction to the long documents. The term co-occurrences in these documents should yield information that is applicable to the rest of the corpus.

If, on the other hand, a document is long because it was generated by an unnatural or irrelevant model (e.g., spam), then the co-occurrence patterns of terms in the document are not as useful. However, these co-occurrence patterns could be highly-represented in the top-ranked singular vectors, which is bad.

2) The documents would have more of a uniform influence on the singular vectors. The effects of long pieces of spam would be greatly diminished, but the positive influence of long documents which happen to contain large amounts of material considered relevant to many users would also be diminished.

If we re-consider the matrix B from problem 1, it's normalized version has the following SVD:

$$B' = \begin{bmatrix} -0.021 & 0.999 & 0.042 & 0.000 & 0.000 \\ -0.962 & -0.031 & 0.271 & 0.000 & 0.000 \\ -0.272 & 0.035 & -0.962 & 0.000 & 0.000 \end{bmatrix} \times \begin{bmatrix} 1.628 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.419 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.579 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{bmatrix} \times \ldots$$

$$\begin{bmatrix} -0.417 & 0.507 & 0.256 & -0.710 & 0.000 \\ -0.452 & 0.462 & 0.298 & 0.702 & 0.018 \\ 0.420 & 0.505 & -0.246 & 0.025 & 0.713 \\ 0.398 & 0.525 & -0.269 & 0.043 & -0.701 \\ -0.536 & 0.002 & -0.844 & 0.012 & 0.022 \end{bmatrix}$$

As we can see, the effect of the “outlying” document has been greatly reduced. The most significant singular vectors roughly correspond to the term-1 and term-2 axes, which is good.