CS 630 Notes: Lecture 17
Lecturer: Lillian Lee

Notes by Matt Connolly, David Lin, and Danish Mujeeb

April 4^{th}, 2006

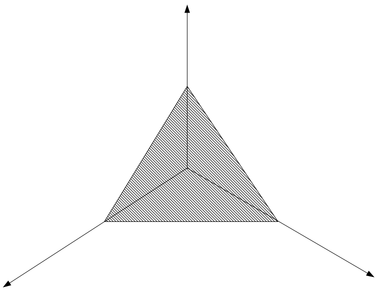

Figure 1: Region formed by vectors in $\mathcal{A}$

1 Review

Recall from last time that we were working with a $m \times n$ term-document matrix $D : \begin{bmatrix} \uparrow & & & & \uparrow \\ \vec{d}^{(1)} & . & . & . & \vec{d}^{(n)} \\ \downarrow & & & & \downarrow \end{bmatrix}$.

We want to characterize the structure of D succinctly, but how?

We first looked at $rank(D)$: the dimensionality of the span of the $\vec{d}^{(i)}$s (where the span represents all possible linear combinations of the $\vec{d}^{(i)}$s). Remember that we can use the expression $\sum_{i=1}^{n} \alpha_i \vec{d}^{(i)} = D\vec{\alpha}$, $\vec{\alpha} \in \Re^n$; D in this case is acting as an operator on the combination coefficients.

Last time, we considered what happens if we drew from only a portion of $\Re^n$: $\mathcal{A} = \{\vec{\alpha} \in \Re^n : \alpha_i \geq 0, \sum \alpha_i = 1\}$. For example: $D : \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}$; this means that $D : \Re^3 \rightarrow \Re^2$ and $\mathcal{A}$ is the region depicted in Figure 1. Then we can look at three different transformations $D'\mathcal{A}$, $D''\mathcal{A}$, and $D'''\mathcal{A}$ which correspond to convex hulls for the column vectors of the respective term-document matrices

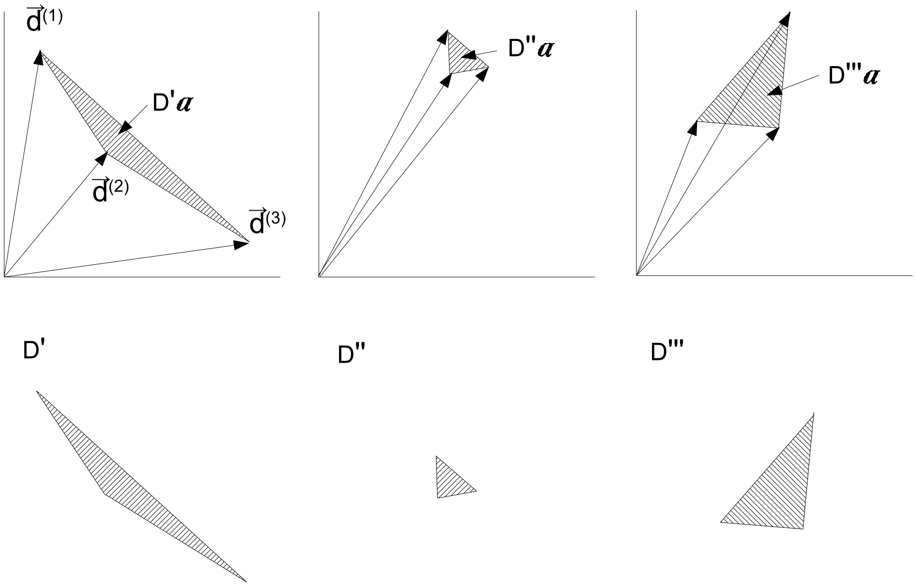

Figure 2: Example transformations of $\mathscr{a}$

D', D'', and D''' due to the special "fractional assignment" nature of $\mathscr{a}$ (see Figure 2). What factor can we use to distinguish these three cases? Shape or area or size don't appear to work; each transformation is substantively different, and yet D'' and D''' in particular intuitively appear to be similar.

2 Applying different subsets of $\Re^n$

2.1 First attempt

So far we've considered all linear combinations, and then a subset of $\Re^n$. Now we'll try looking at a different subset. How about $\mathscr{a}_1 \supseteq \mathscr{a}$, with $\mathscr{a}_1 = \{\vec{\alpha} \in \Re^n : \sum |\alpha_i| = 1\} = \{\vec{\alpha} \in \Re^n : ||\vec{\alpha}||_1 = 1\}$?

Figure 3 shows $\mathscr{a}$ and $\mathscr{a}_1$ for the case of $n = 2$ (two-dimensional space).

In three dimensions: $\mathscr{a}_1$ is an octahedron with six vertices at $(\pm 1, 0, 0)$, $(0, \pm 1, 0)$, $(0, 0, \pm 1)$.

Observe that $D\mathscr{a}_1$ forms the convex hull of the $\pm \vec{d}^{(i)}$s. Thus our projections and corresponding "regions" appear as in Figure 4.

These projections differ in their shape: the first one, $D'\mathscr{a}_1$, is "fat" – which makes sense because the document vectors point in numerous disparate directions – while the others have more directionality.

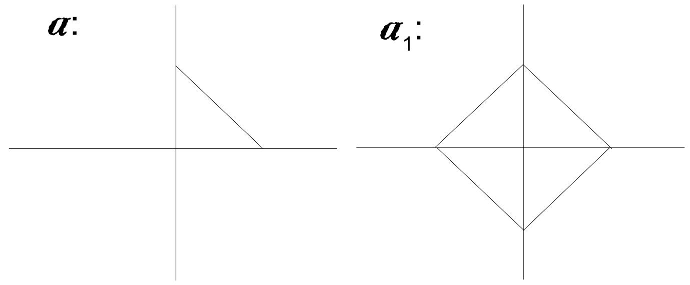

Figure 3: Regions defined by a and a_1

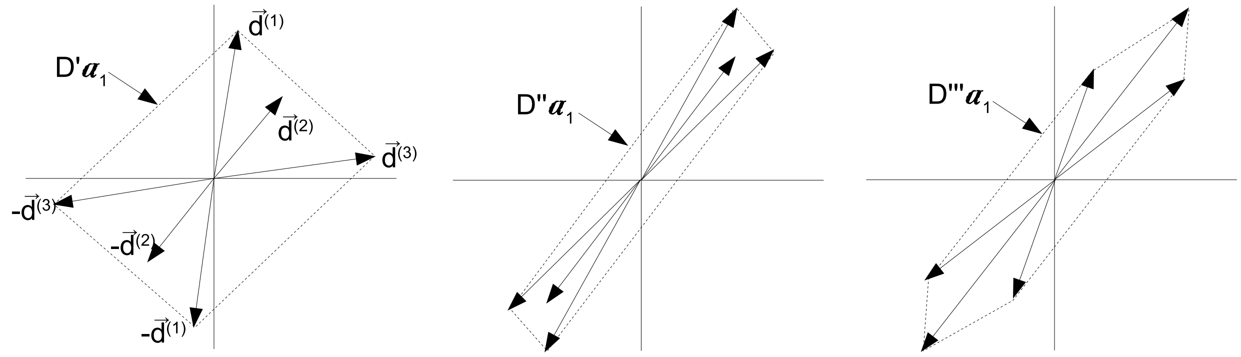

Figure 4: Transformations for a_1

This gives us the notion of "directed area". However, there is still something not quite right about this description. For one thing, these polyhedra are hard to describe; for another, some of them are degenerate.

2.2 Second attempt

Let's try a different but related $\Re^n$ subset in hopes of a better result. Let $a_2 = \{\vec{\alpha} \in \Re^n : ||\vec{\alpha}||_2 = 1\}$ (so we are looking at $\vec{\alpha}$ normalized to the 2-norm as opposed to the 1-norm). a_2 in 2-d space is described in Figure 5.

Now our transforms will be mapped to hyperellipses that contain the vertices of the convex hulls we had previously (see Figure 6). (The circular or hyperspherical a_2 region is providing a smoothing effect in this case.) The shapes provide a good characterization of the corpus.

All we need now for a description is a characterization of the axes of these hyperellipsoids. These hyperellipsoids can be described by the direction and lengths of their semimajor axes using the following terminology:

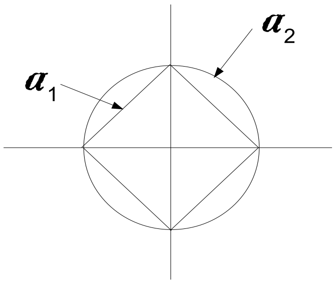

Figure 5: a_2 in 2D

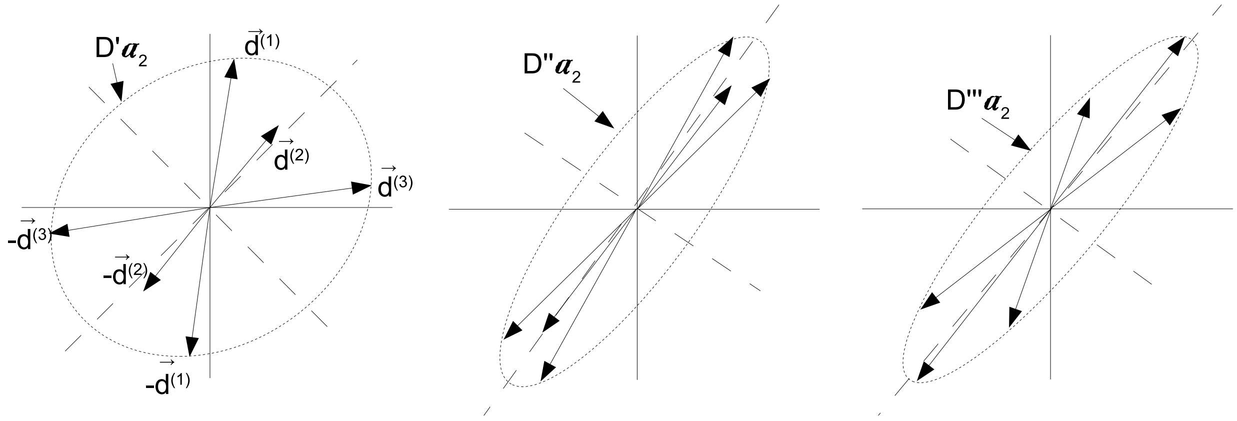

Figure 6: Transformations of a_2

- $\vec{u}^{(1)}, ..., \vec{u}^{(r)} \in \Re^m$ orthonormal: describe the axes of an ellipsoid
- $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$: lengths along the axes
- If the σ_l values are distinct, then the $\vec{u}^{(l)}$ values are distinct up to sign.
- For each $\vec{u}^{(l)}$ there is a corresponding unit vector $\vec{v}^{(l)} \in \Re^n$ such that $D\vec{v}^{(l)} = \sigma_l \cdot \vec{u}^{(l)}$.

Now, how do we obtain these ellipsoid values? One way would be to go through some rather arduous calculations. Luckily, though, there exist algorithms that will produce three matrices (the SVD matrices):

$$D = \begin{bmatrix} \uparrow & & & & \uparrow \\ \vec{u}^{(1)} & . & . & . & \vec{u}^{(r)} \\ \downarrow & & & & \downarrow \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & . & 0 & \\ & & . & \\ & 0 & & . & \\ & & & & \sigma_r \end{bmatrix} \begin{bmatrix} \leftarrow & \vec{v}^{(1)} & \rightarrow \\ & . & \\ & . & \\ \leftarrow & \vec{v}^{(r)} & \rightarrow \end{bmatrix}$$

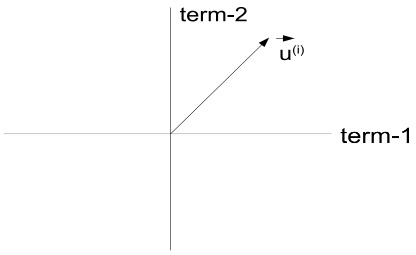

Figure 7: A $\vec{u}^{(l)}$ vector in 2-term space, perhaps showing that Term 1 and Term 2 tend to co-occur in documents.

The $\vec{v}^{(l)}$'s are the preimages of the $\sigma_l \vec{u}^{(l)}$'s and map onto the axes. They also form an orthonormal basis for the row (term) space; the i^{th} column of the $\vec{v}$ matrix forms the coordinates for $\vec{d}^{(i)}$ in the $\vec{u}^{(l)}$ basis. The $\vec{u}^{(l)}$ are known as left singular vectors, the $\vec{v}^{(l)}$ are termed right singular vectors, and the σ_l's are the singular values.

2.3 Relating to IR

But what does all of this have to do with information retrieval and documents? What do these vectors and matrices actually represent?

Are the $\vec{u}^{(l)}$'s topics? (Wouldn't that be nice?) Well, the $\vec{u}^{(l)}$'s are the axes of the ellipsoid; in particular, $\vec{u}^{(1)}$ is the longest axis (see Figure 7). And, if we think of $\vec{u}^{(l)}$ as a document, then it's one that contains Term 1 and Term 2 equally often.

So one interpretation of these vectors is that they indicate "co-occurrence" patterns, or how often multiple terms may be found together.

It may be tempting to think of these patterns as topics. However, this interpretation runs into problems when we consider the fact that the left singular vectors must be orthogonal, which in vector terms means 'equally, and very, far apart'. In contrast, topics can be 'different degrees apart': for instance, we could have a corpus in which the documents appear to correspond to the topics "politics", "dogs", and "cats", where "dogs" seems closer to "cats" than to "politics".

One thing often said in the literature is that the $\vec{u}^{(l)}$ form the directions of "greatest variation". Of course, in the case of a given ellipse, there are different groups of vectors that could produce its shape and directionality (see Figure 8). So this interpretation is really talking about the greatest variation *away from zero*, not variability between vectors.

SVD gives us

$$D = U\Sigma V^T$$

What makes this r-dimensional U-basis special amongst the infinite number of bases for the same subspace (assuming $r > 1$)? As it turns out, it gives us a ranking function: the $\vec{u}^{(l)}$'s are ranked by the σ_l's!

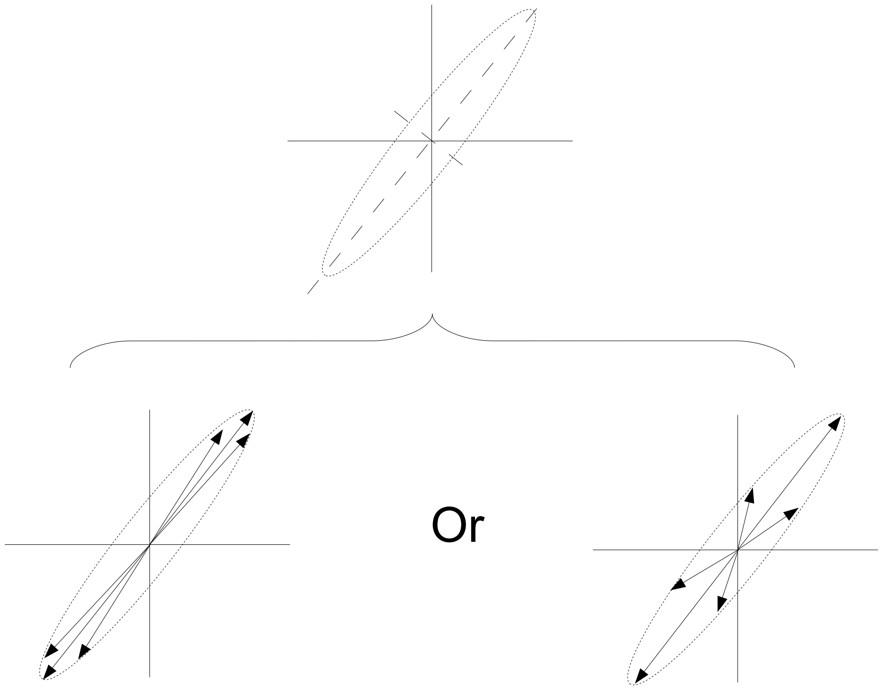

Figure 8: Elliptical projection and example source vector groups

2.4 Latent Semantic Indexing

Suppose we project D into a space spanned by the "top k" $\vec{u}^{(l)}$'s. This preserves as much ellipse area (surface area) as possible by taking the longest axes. According to a theorem by Eckart and Young (1936), this gives us the best rank-k approximation to D (in a mathematical sense).

In an IR sense, though, it's hard to come up with a definitive interpretation that says that doing this is helpful. But we can hope!

One hope is that, if we're lucky, the $\vec{u}^{(l)}$'s do tell us about important co-occurence patterns. For example, given the terms "humpty" and "dumpty", which almost always occur together, we can hope that a single $\vec{u}^{(1)}$ vector represents this co-occurrence of two terms (e.g. as depicted in Figure 7). Thus we've saved a description dimension, since before we would have represented each term individually.

Another hope is that what gets left over (the low-ranking $\vec{u}^{(l)}$'s) corresponds to some sort of "noise" in our system (e.g., synonyms).

Papadimitriou et al. posited that for "distinct vocabulary" topic models, LSI can derive the underlying structure. But the problem is that D looks something like $\begin{bmatrix} [\ \ ] & & \\ & & [\ \ ] \\ & [\ \ ] & \end{bmatrix}$, with each sub-block of non-zero entries indicating which topic model generated that group of column vectors. (That is, block matrices that possess obvious sub-block structures have a structure difficult for any algorithm to miss.)

We will discuss LSI further in the next lecture.

3 Questions

1. Recall that we decided to characterize the structure of our term-document matrix D succinctly by using linear combinations of the $\vec{d}^{(i)}$s with the combination coefficients drawn from $\{\vec{\alpha} \in \Re^n : ||\alpha||_2 = 1\}$, where n is the number of documents. These linear combinations provided us with a hyperellipsoid non-uniquely described by the following parameters:

    - $\vec{u}^{(1)}, ..., \vec{u}^{(r)} \in \Re^m$: orthonormal vectors that describe the axes of ellipse. m is the number of terms in our term-document matrix, and r is its rank.
    - $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$: lengths along the axes

    Consider a document space where the document vectors (only two terms) are arranged as in Figure 9 :

    We see that there are basically two axes that would capture most of the variation of our document space. We could take the length of the hyperellipsoid's "radius" along each axis. However, we would find that the hyperellipsoid would have a major axis defined by the lone document vector on the left side of the document space, since that vector has a longer length than any other. Would there be an alternative and more informative number we could associate with each axis of the hyperellipsoid that would better characterize this document corpus?

    Answer: In calculating a better measure of the importance of each axis, we could also take into account the number of documents along each axis. We could calculate the importance of each axis by calculating the total length of the projections of all documents in the corpus onto that axis. This measure would tell us which direction most of the "mass" of the document space is concentrated along. With the example document corpus, we would then find that the major axis of the hyperellipsoid should lie along the clustered document vectors on the right side of the document space.

2. Consider a case where we have two document vectors that we want to compare to a fixed query:

$$\vec{q} = (3, 2, 3)$$
$$\vec{d}^{(1)} = (1, 5, 3)$$
$$\vec{d}^{(2)} = (4, 2, 3)$$

    Thus we incorporate these vectors into our term-document matrix: $D = \begin{bmatrix} 3 & 1 & 4 \\ 2 & 5 & 2 \\ 3 & 3 & 3 \end{bmatrix}$.

    We have calculated the SVD of D for you using MatLAB:

$$U = \begin{bmatrix} -0.5 & 0.7 & -0.5 \\ -0.6 & -0.7 & -0.4 \\ -0.6 & 0.1 & 0.8 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 8.7 & 0 & 0 \\ 0 & 3.2 & 0 \\ 0 & 0 & 0.3 \end{bmatrix}$$

$$V = \begin{bmatrix} -0.5 & 0.3 & 0.8 \\ -0.6 & -0.8 & -0.1 \\ -0.6 & 0.5 & -0.6 \end{bmatrix}$$

Recall that in order to obtain the best rank-k approximation of D, we simply set $\sigma_{k+1}, ..., \sigma_n = 0$. Calculate the similarity between $\vec{d}^{(1)}$ and $\vec{q}$ and between $\vec{d}^{(2)}$ and $\vec{q}$ for $k = 1, ..., 3$. Does the ranking of the documents with respect to the query change depending on what rank approximation we use? In this simple case, we will calculate similarity between the document and query by using the dot product of the two vectors.

Answer: For $k = 1$,

$$\Sigma_1 = \begin{bmatrix} 8.7 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, D_1 = \begin{bmatrix} 2.4 & 2.8 & 2.7 \\ 2.8 & 3.3 & 3.1 \\ 2.7 & 3.2 & 3.0 \end{bmatrix}$$

$$q \cdot \vec{d}^{(1)} = 24.5$$

$$q \cdot \vec{d}^{(2)} = 23.3$$

For $k = 2$,

$$\Sigma_2 = \begin{bmatrix} 8.7 & 0 & 0 \\ 0 & 3.2 & 0 \\ 0 & 0 & 0 \end{bmatrix}, D_2 = \begin{bmatrix} 3.1 & 1.0 & 4.0 \\ 2.1 & 5.0 & 2.0 \\ 2.8 & 3.0 & 3.2 \end{bmatrix}$$

$$q \cdot \vec{d}^{(1)} = 22.0$$

$$q \cdot \vec{d}^{(2)} = 25.0$$

For $k = 3$,

$$\Sigma = \begin{bmatrix} 8.7 & 0 & 0 \\ 0 & 3.2 & 0 \\ 0 & 0 & 0.3 \end{bmatrix}, D = \begin{bmatrix} 3.0 & 1.0 & 4.0 \\ 2.0 & 5.0 & 2.0 \\ 3.0 & 3.0 & 3.0 \end{bmatrix}$$

$$q \cdot \vec{d}^{(1)} = 22.0$$

$$q \cdot \vec{d^{(2)}} = 25.0$$

Thus we see that using a reduced representation, rank-1 approximation, of D gives us a different ranking of the documents with respect to the query compared to that which we obtained from D.

Also note that the column vectors of D_2 are quite similar to those in the original term-document matrix D, whereas the document vectors in D_1 are quite different from the originals.

3. Recall that we considered shape and area as potentially informative properties of our mappings of different polygons and polyhedra before settling on the directionality and lengths of the axes of an ellips(oid).

   In two-dimensional space, another descriptive property of an ellipse is its eccentricity, defined as $e = \sqrt{1 - \frac{b^2}{a^2}}$, where a is the length of the semimajor axis and b the length of the semiminor axis. A perfect circle has an eccentricity of 0. Thus the eccentricity can provide us with a one-number indication of an ellipse's overall shape. But is this useful to us?

   - In a two-dimensional term space, measuring the eccentricity of the ellipse bounding our document vectors (and negatives thereof) can describe the similarity of documents in a corpus. *For this purpose consider the document lengths to be normalized* and negative term values to be acceptable. A set of documents with similar topics or terms will form vectors that point along the same, or similar, angles in the term space. The resultant ellipse will be highly elongated and possess a proportionally higher eccentricity. On the other hand, a corpus with documents of highly distributed subjects will produce a more rounded ellipse with a much lower eccentricity. (Note, though, that a rounded ellipse may also be produced by other distributions, provided that there is some disparity among topics in the overall corpus.) In this application, then, eccentricity is sometimes similar to an IDF term in the VSM.
   - However, while the eccentricity informs us about *shape*, it provides no information about *direction*. Two sets of documents with similar distributions, but very different topic concentrations, will possess similar eccentricities. In this sense the measurement of e is less informative than other methods.
   - (Open-ended) Can e help us with more complex (i.e., higher-dimensional) document models? In most cases, eccentricity is confined to a two-dimensional representation of ellipses; in other applications, typically the eccentricity of a cross-section is measured instead. Would this still be useful? Is it possible to derive an n-dimensional equivalent to eccentricity that would give us the same sort of "shape" information?

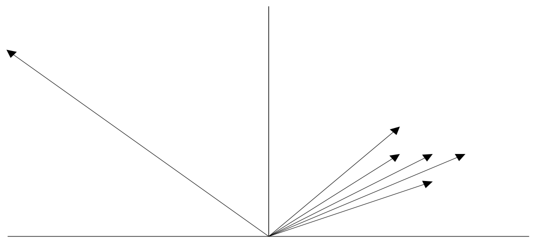

Figure 9: Figure for Question 1