March 25, 2009

Lecture 26. Singular Value Decomposition (SVD)

Form

$A = U\Sigma V^T$

U and V have orthonormal columns and Σ is diagonal

Transformation

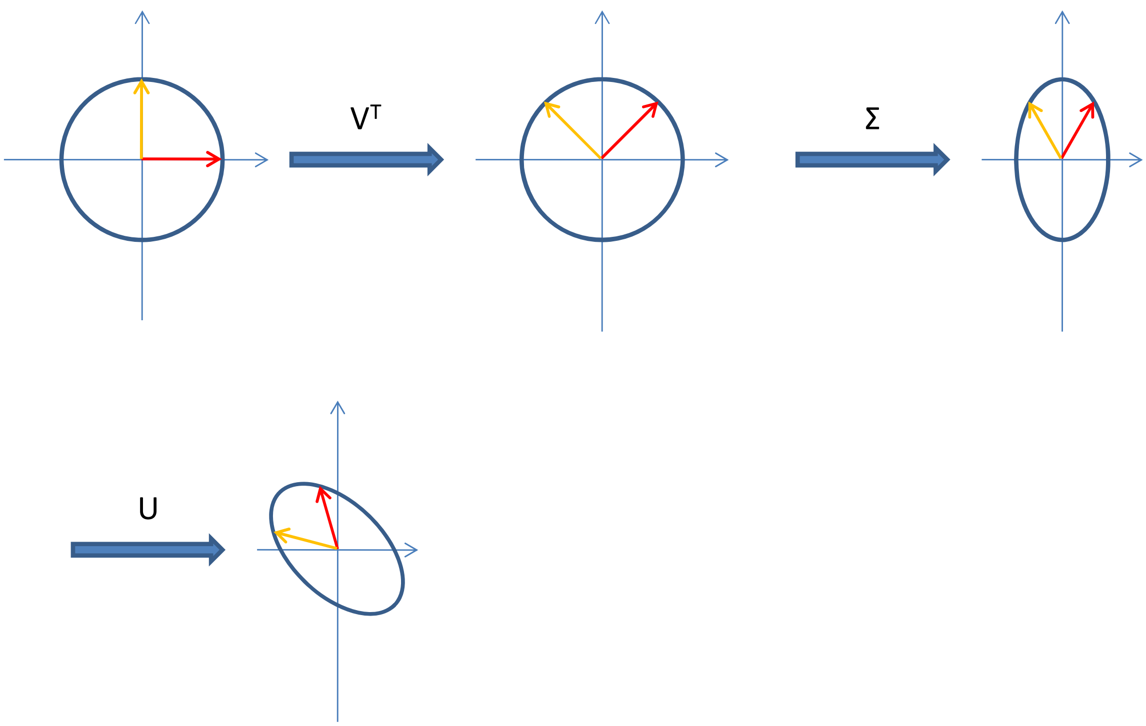


Figure 1: Demonstrating the steps of SVD components' transformation

Reasons to be interested in it:

1) V^T stationary probability of random walk
2) Finding subspace that minimizes sum of square of distance to set of points

Fitting data

The task is to find the subspace that minimizes the sum of squares of distance to a set of points. Our subspace will be the line of the form $y = f(x) = ax + b$.

This is useful when dealing with word vectors. A word vector is a data structure in which vectors represent documents and elements in the vectors correspond to the frequencies of words in the documents.

When doing linear regression by minimizing mean squares errors, we need to minimize $\sum_{i=1}^{n}\left[y_i-(ax_i+b)\right]^2$

This can be solved by differentiating $\frac{\partial}{\partial a}$ and $\frac{\partial}{\partial b}$.

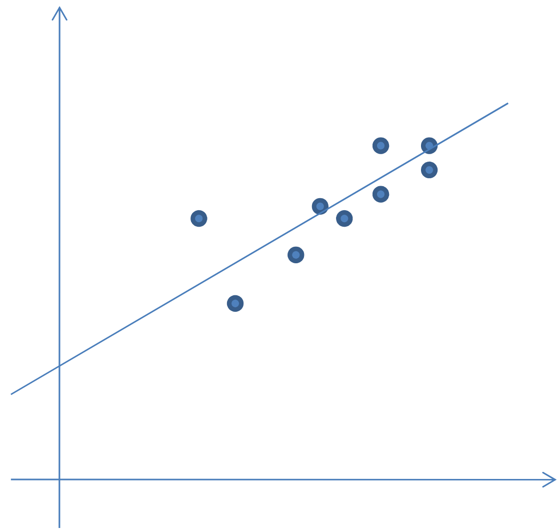


Figure 2: Fitting by minimizing square errors

However, our problem is slightly different than conventional linear regression. We force $f(x)$ to go through the origin and measure the distance perpendicular to the line rather than vertically.

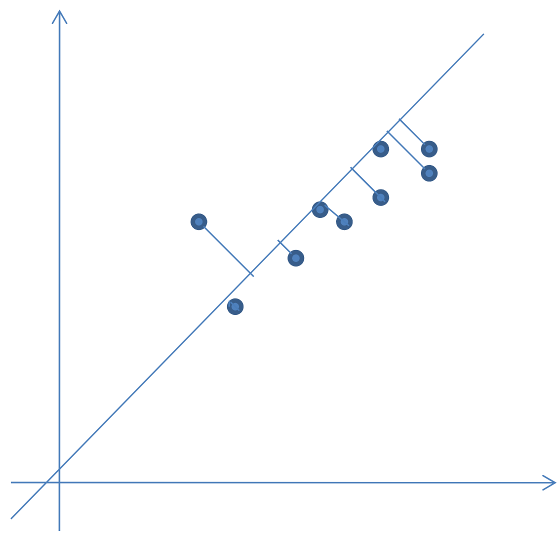


Figure 3: Best fit subspace by minimizing distance

$\{(x_i, y_i) | 1 \le i \le n\}$

To minimize absolute values or square of norms have different implications. This could be a potential homework problem. The key lies in the treatment of the larger errors relative to the smaller ones.

To find the best fit subspace:

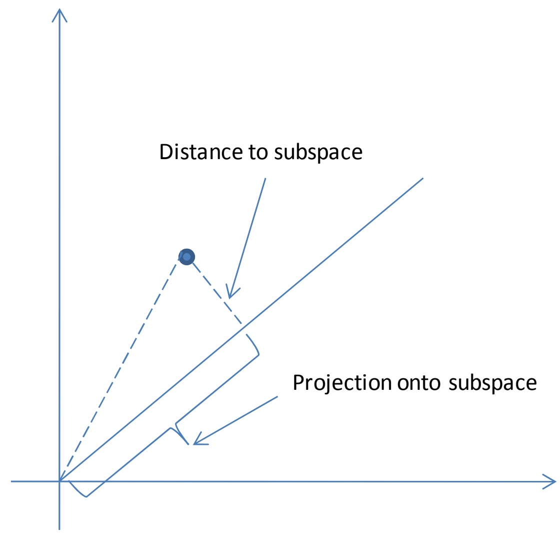


Figure 4: Distance and Projection

Let d_i be the perpendicular distance from the subspace to the i^{th} point, and l_i be the length of the perpendicular projection onto the subspace.

$$\sum_{i=1}^{n}(x_i^2 + y_i^2) = \sum_{i=1}^{n} d_i^2 + \sum_{i=1}^{n} l_i^2$$

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}(x_i^2 + y_i^2) - \sum_{i=1}^{n} l_i^2$$

Since $\sum_{i=1}^{n}(x_i^2 + y_i^2)$ is a constant, minimizing $\sum_{i=1}^{n} d_i^2$ is the same as maximizing $\sum_{i=1}^{n} l_i^2$.

<u>Similarity Matrix of documents</u>

In the word vector model, assume you have a corpus in which there are documents on three different subjects.

Let *A* be the similarity matrix in which the columns and rows correspond to documents and the ij^{th} entry is a number between zero and one corresponding to the similarity between the i^{th} document and the j^{th} one. All of the comparisons between documents on the same subject will have essentially the same high value. All of the comparisons between documents on different subjects will have a value of about zero.

$$\begin{pmatrix} \square_1 & 0 & 0 \\ 0 & \square_2 & 0 \\ 0 & 0 & \square_3 \end{pmatrix}$$

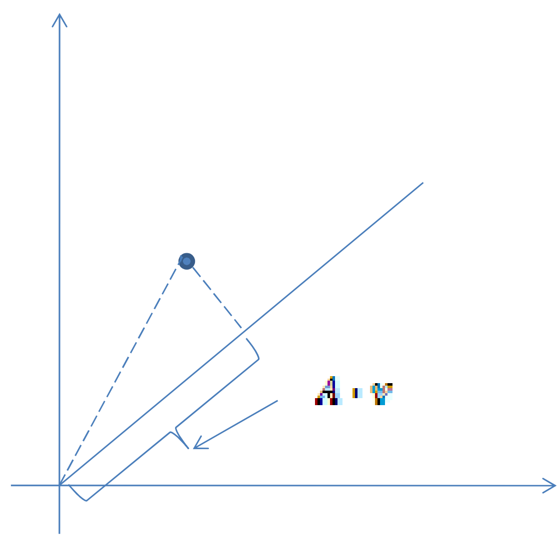
If you consider each row as a point in space, you can create a best fit line through the data. All of the points will cluster around three points, one of which corresponds to each subject. The best fit "line" through the origin for this data will help us pull out the cluster of documents.

Then for example, we would have a vector $\vec{v_1} = (1,1,1,\ldots,0,0,0,\ldots)$, that runs through the best fit line and $\vec{v_2} \perp \vec{v_1}$ that minimizes the sum of squares of the d_i.

This is exactly what SVD, singular value decomposition does.

<u>Minimizing distances and SVD</u>

Let A be an n x d matrix in which the rows are points in d dimensions. Let v be the direction of the best fit line through the origin and normalize v such that $|v| = 1$.



The length of the projection of i^{th} row is $|a_i \cdot v|$

$$\sum_{i=1}^{n} l_i^2 = \sum_{i=1}^{n} |a_i \cdot v|^2 = |Av|^2$$

Find $v_1 = \arg\max |Av|^2, \quad |v| = 1$ (constraint)

(i.e. the unit length vector v_1 such that $|Av|^2$ is maximized)

v_1 is the first right singular vector.

$\sigma_1 = |Av_1|^2$ is the first singular value.

$v_2 = \arg\max |Av_1|^2, \quad |v| = 1, \quad v_2 \perp v_1$ (constraints)

v_2 is the second right singular vector.

$u_1 = \frac{1}{\sigma_1} Av_i$ is the i^{th} left singular vector.

This process gives us u_i, σ_i, v_i which form U, Σ and V matrices in the singular value decomposition.

<u>Proof of Claim</u>

To validate this, we need to prove two things:

1) The u_i vectors are orthogonal
2) $A = \sum \sigma_i u_i v_i^T$

As it is trivial to see v_i are orthonomal.

Proof: We prove them by induction. For the base case, where it's one dimension, 1) is trivially true. Only 1 u_i exists. Then, we need to prove 2). In the next lecture, we will continue the proof, using the following lemma:

Lemma: To prove A=B, it suffices to show that for all v, Av=Bv.

(END OF LECTURE)