SINGULAR VALUE DECOMPOSITION & SUBSET OF BEST FIT (11/13/2020)

Review:
We have been working with SVD for quite a few days. All we need to recall here is that given an $n \times m$ matrix A, we see that $A = U\Sigma V^T$, or $A = U_r \Sigma_r V_r^T$. Furthermore, U and V are orthonormal. Finally, the bases of the subspaces of A can be determined by columns of U and V.

SVD as Change of Basis:
We can also think about the matrix equation $A = U\Sigma V^T$ as a change of basis. Recall for $m \times n$ matrix A, we have U is $m \times m$, Σ is $m \times n$, and V is $n \times n$. Think back to our original picture of the four fundamental subspaces! The matrix A allowed a transportation from $\mathbb{R}^n$ to $\mathbb{R}^m$, and $C(A^T)$, $N(A)$ existed orthogonally in $\mathbb{R}^n$ while $C(A)$, $N(A^T)$ existed orthogonally in $\mathbb{R}^m$. As we know $A\vec{v}_i = \sigma_i \vec{u}_i$, we can determine that:

- $\vec{v}_1, \vec{v}_2, \ldots \vec{v}_r$ basis of $C(A^T)$
- $\vec{v}_{r+1}, \vec{v}_{r+2}, \ldots \vec{v}_n$ basis of $N(A)$
- $\sigma_1\vec{u}_1, \sigma_2\vec{v}_2, \ldots \sigma_r\vec{v}_r$ basis of $C(A)$
- $\sigma_1\vec{u}_{r+1}, \sigma_2\vec{v}_{r+2}, \ldots \sigma_r\vec{v}_n$ basis of $N(A^T)$

Using this information, we can think about each individual matrix V, Σ, and U.

- The matrix V^T (or V^{-1}) maps $\vec{v}_1, \vec{v}_2, \ldots \vec{v}_n$ to $\vec{e}_1, \vec{e}_2, \ldots \vec{e}_n$, or the standard basis vectors! So, V is an orthogonal transformation (rotation/reflection) to a change of basis.
- Now, from here the matrix Σ scales $\vec{e}_1, \vec{e}_2, \ldots \vec{e}_n$ to $\sigma_1\vec{e}_1, \sigma_2\vec{e}_2, \ldots \sigma_n\vec{e}_n$. This is effectively a stretch of each coordinate!
- Next, from here the matrix U is another rotation/reflection. It sends our standard basis $\vec{e}_1, \vec{e}_2, \ldots \vec{e}_n$ to $\vec{u}_1, \vec{u}_2, \ldots \vec{u}_n$. Note that this reverts multiplying by V^T and U.

So, with singular value decomposition, we can see exactly what multiplying by matrix A does – a change of basis, then a stretching of the basis, and another change of basis.

Best-fit Subspace
There are a lot of very interesting applications of the SVD in mathematics and statistics. We will go over three of these: best-fit subspace, low rank matrix approximation, and principal component analysis. Today, per the title, we are going to investigate best-fit subspace.
Suppose we have a lot of different points in $\mathbb{R}^n$. We would really like to know what the "best-fit" m-dimensional subspace is, where $m < n$. Why would we like to do this? Well, suppose we have tons of data in 3D or higher dimensions that is difficult to visualize on a page of a screen. Well, we could represent it with the highest fidelity using the best-fit 2-dimensional subspace!
So, given vectors $\vec{x}_1, \ldots \vec{x}_m \in \mathbb{R}^n$, we aim to find an orthogonal set of vectors $\{\vec{v}_1, \ldots \vec{v}_k\}$ spanning a k-dimensional subspace $V \subseteq \mathbb{R}^n$ such that:

- The $\sum_{i=1}^m ||\vec{e}_i||^2$, or the error squared, is minimized. This is from least squares approximation.
- The $\sum_{i=1}^m ||\vec{p}_i||^2$, or the length of all the projections, is minimized.

Note that this concept of the error squared and the length of the projections is taken from least squares approximation. Furthermore, note that is these conditions above are equivalent! This can be derived from the equation (think Pythagorean Theorem):

$$\sum_{i=1}^m ||\vec{x}_i||^2 = \sum_{i=1}^m \left(||\vec{p}_i||^2 + ||\vec{e}_i||^2\right)$$

We basically want to find an orthonormal set $\{\vec{v}_1, \ldots \vec{v}_k\}$ such that for $1 \leq i \leq n$ we find:

$$\vec{x}_i \approx \vec{p}_i = \text{proj}_{\vec{v}_1}\vec{x}_i + \ldots \text{proj}_{\vec{v}_k}\vec{x}_i = (\vec{x}_i^T\vec{v}_1)\vec{v}_1 + \ldots (\vec{x}_i^T\vec{v}_k)\vec{v}_k$$

A Set of Quick Propositions:
Before we continue, we need to prove a few propositions. If Λ is a positive semi-definite diagonal matrix, then the unit vector $\vec{y}$ maximizing $\vec{y}^T\Lambda\vec{y}$ is a standard basis vector $\vec{e}_i$ where λ_i is the largest value on the diagonal of Λ. This is relatively obvious – we can prove it using the fact $\vec{e}_i^T\Lambda\vec{e}_i = \lambda_i$. Note furthermore that as λ_1 is the largest value on the diagonal for any positive semi-definite diagonal matrix, we conclude that the maximizing vector must be $\vec{y} = \vec{e}_1$.

Our next proposition is if A is a matrix with an SVD $A = U\Sigma V^T$, then A^TA has a spectral decomposition $A^TA = Q\Lambda Q^T$. Simply let $Q = V$ and $\Lambda = \Sigma^T\Sigma$. We see that:

$$A^TA = (V\Sigma^TU^T)(U\Sigma V^T) = V\Sigma^T\Sigma V^T$$

Finally, we have that if A is a matrix with an SVD $A = U\Sigma V^T$, and thus $A^TA = Q\Lambda Q^T = V\Sigma^T\Sigma V^T$, then the unit vector that maximizes $||A\vec{v}||$ is simply $\vec{v}_1$, the first singular vector. The proof is as follows:

$$||A\vec{v}||^2 = (A\vec{v})^T(A\vec{v}) = \vec{v}^TV\Sigma^T\Sigma V^T\vec{v} = \vec{y}^T(\Sigma^T\Sigma V^T)\vec{y}$$

From our last equation, we know that this value above is maximized when $\vec{y} = \vec{e}_1$. So $\vec{e}_1 = V^T\vec{v}$ by multiplying by V, we see that:

$$\vec{e}_1 = V^T\vec{v} \implies V\vec{e}_1 = \vec{v} \implies \vec{v}_1 = \vec{v}$$

So, we are done! Note these propositions are really just quantifications of a lot of properties inherent in SVD.

Wrapping it All Up
So now, we have the tools to get the best-fit subspace! Let's consider solely the case in which $k = 1$. Then, for a single $\vec{v}$ we simply want to maximize:

$$\sum_{i=1}^{m} \text{proj}_{\vec{v}}\vec{x_i} = \sum_{i=1}^{m} ||(\vec{x_i}^T\vec{v})\vec{v}||^2 = \sum_{i=1}^{m}(\vec{x_i}^T\vec{v})^2$$

We see that if we form a matrix A the rows of which are $\vec{x_1}^T \dots \vec{x_m}^T$, then it follows that the above is equivalent to $||A\vec{v}||$. Thus, it follows quite simply that we can make $A = U\Sigma V^T$, and then let $\vec{v} = \vec{v}_1$, the first column of V. The one-dimensional subspace of best fit is simply spanned by $\vec{v}_1$. Thus, we are finished, and our subspace is maximized! If we would like to approximate higher-dimensional subspaces, we simply continue to take columns of V! The k-dimensional subspace of best fit will be $\{\vec{v}_1, \dots \vec{v}_k\}$.