Singular Value Decomposition

Psych 267/CS 348D/EE 365

Prof. David J. Heeger

September 15, 1998

This handout is a review of some basic concepts in linear algebra. For a detailed introduction, consult a linear algebra text. Linear Algebra and its Applications by Gilbert Strang (Harcourt, Brace, Jovanovich, 1988) is excellent.

1 Singular Value Decomposition and the Four Fundamental Subspaces

The SVD decomposes a matrix into the product of the three components:

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^t,$$

where $\mathbf{V}^t$ means transpose. Here, $\mathbf{A}$ is the original NxM matrix, $\mathbf{U}$ is an NxN orthonormal matrix, $\mathbf{V}$ is an MxM orthonormal matrix, and $\mathbf{S}$ is an NxM matrix with non-zero elements only along the main diagonal. The number of non-zero elements in $\mathbf{S}$ is (at most) the lesser of M and N. The rank of a matrix is equal to the number of non-zero singular values.

We think of $\mathbf{A}$ as a linear transform, $\mathbf{A}\mathbf{x} = \mathbf{b}$, that transforms $\mathbf{x}$ (M-dimensional vectors) into $\mathbf{b}$ (N-dimensional vectors). If the matrix $\mathbf{A}$ is singular then there is some collection of $\mathbf{x}_i$ (some subspace of M-space) that is mapped to zero, $\mathbf{A}\mathbf{x}_i = \mathbf{0}$. This is called the "row nullspace" of $\mathbf{A}$. There is also a subspace of M-space that can be reached by $\mathbf{A}$, i.e., the set of vectors $\mathbf{b}_i$ for which there is some $\mathbf{x}$ where: $\mathbf{A}\mathbf{x} = \mathbf{b}_i$. This is called the "column space" of $\mathbf{A}$. The dimension of the column space is called the rank of $\mathbf{A}$.

The SVD explicitly constructs *orthonormal bases* for the row nullspace and column space of $\mathbf{A}$. The columns of $\mathbf{U}$, whose same-numbered elements in $\mathbf{S}$ are non-zero, are an orthonormal set of basis vectors that span the column space of $\mathbf{A}$. The remaining columns of $\mathbf{U}$ span the row nullspace of $\mathbf{A}^t$ (also called the column nullspace of $\mathbf{A}$). The columns of $\mathbf{V}$ (rows of $\mathbf{V}^t$), whose same-numbered elements in $\mathbf{S}$ are zero, are an orthonormal set of vectors that span the row nullspace of $\mathbf{A}$. The remaining columns of $\mathbf{V}$ span the column space of $\mathbf{A}^t$ (also called the row space of $\mathbf{A}$).

Here's a simple example:

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & -1 \end{pmatrix}.$$

In this example:

- The matrix size is 3x3 (M and N are both 3).

- The rank of the matrix is 2.
- $(1, 1, 1)$ is in the row nullspace. So is any scalar multiple of this vector.
- $(1, 1, -1)$ is in the column nullspace. So is any scalar multiple of this vector.
- $(-2, 1, 1)$ and $(0, 1, -1)$ are in the row space. So is any linear combination of these two vectors.
- $(-1, 0, -1)$ and $(-1, 2, 1)$ are in the column space. So is any linear combination of these two vectors.

It is easy to see that $\mathbf{A}$ has rank 2 by noting that you can get the third column of $\mathbf{A}$ by summing the first 2 columns and then multiplying by -1. The vector (1,1,-1) is in the column nullspace. It is called the column nullspace because it takes to columns to zero: $(1, 1, -1)\mathbf{A} = 0$.

The vector $(1, 1, 1)$ is in the row nullspace of A. It is called the row nullspace because it takes to rows to zero: $\mathbf{A}(1, 1, 1)^t = \mathbf{0}$.

How about the column space? Consider all possible combinations of the columns, $\mathbf{Ax}$, coming from all choices of $\mathbf{x}$. Those products form the column space of $\mathbf{A}$. In our example, the column space is a 2-dimensional subspace of 3-space (a tilted plane), linear combinations of: $(-1, 0, -1)$ and $(-1, 2, 1)$. Check that all 3 columns of $\mathbf{A}$ can be written as linear combinations of these two vectors. In general, the column space of a matrix is orthogonal to the column nullspace. This is easy to see in our example: $(1, 1, -1) \cdot (-1, 0, -1) = 0$ and $(1, 1, -1) \cdot (-1, 2, 1) = 0$.

How about the row space? Consider all possible combinations of the rows, $\mathbf{b}^t\mathbf{A}$, coming from all choices of $\mathbf{b}$. Those products form the row space of $\mathbf{A}$. In our example, the row space is another 2-dimensional subspace of 3-space (a different tilted plane), linear combinations of: $(-2, 1, 1)$ and $(0, 1, -1)$. Check that all 3 rows of $\mathbf{A}$ can be written as linear combinations of these two vectors. In general, the row space of a matrix is orthogonal to the row nullspace. In our example: $(1, 1, 1) \cdot (-2, 1, 1) = 0$ and $(1, 1, 1) \cdot (0, 1, -1) = 0$.

2 Linear Systems of Equations

Matrices are a convenient way to solve systems of linear equations, $\mathbf{Ax} = \mathbf{b}$. Consider the same matrix, $\mathbf{A}$, we used above. The product, $\mathbf{Ax}$ is always a combination of the columns of $\mathbf{A}$:

$$\mathbf{Ax} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + x_2 \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}.$$

To solve $\mathbf{Ax} = \mathbf{b}$ is to find a combination of the columns that gives $\mathbf{b}$. We consider all possible combinations $\mathbf{Ax}$, coming from all choices of $\mathbf{x}$. Those products form the column space of $\mathbf{A}$. In the example, the columns lie in 3-dimensional space, but their combinations fill out a plane (the matrix has rank 2). The plane goes through the origin since one of the combinations has weights $x_0 = x_1 = x_2 = 0$. Some (in fact, most) vectors $\mathbf{b}$ do not lie on the plane, and for them, $\mathbf{Ax} = \mathbf{b}$, can not be solved exactly. The system, $\mathbf{Ax} = \mathbf{b}$, has an exact solution only when the right side $\mathbf{b}$ is in the column space of $\mathbf{A}$.

The vector $\mathbf{b} = (2,3,4)^t$ is not in the column space of $\mathbf{A}$ so for this choice of $\mathbf{b}$ there is no $\mathbf{x}$ that satisfies $\mathbf{Ax} = \mathbf{b}$. The vector, $\mathbf{b} = (2,3,5)^t$, on the other hand, does lie in the plane (spanned by the columns of $\mathbf{A}$) so there is a solution. In particular, $\mathbf{x} = (5,3,0)^t$ is a solution (try it). However, there are other solutions as well; $\mathbf{x} = (6,4,1)^t$ will work. In fact, we can take $\mathbf{x} = (5,3,0)$ and add any scalar multiple of $\mathbf{y} = (1,1,1)$ since (1,1,1) is in the row nullspace of $\mathbf{A}$. For any scalar constant, c, we can write:

$$\mathbf{A}(\mathbf{x} + c\mathbf{y}) = \mathbf{Ax} + \mathbf{A}(c\mathbf{y}) = \mathbf{Ax} + c\mathbf{Ay} = \mathbf{Ax}$$

If an NxN matrix $\mathbf{A}$ has linearly independent columns then:

1. The row nullspace contains only the point $\mathbf{0}$.

2. The solution to $\mathbf{Ax} = \mathbf{b}$ (if there is one) is unique.

3. The rank of $\mathbf{A}$ is N.

In general any two solutions to $\mathbf{Ax} = \mathbf{b}$ differ by a vector in the row nullspace.

3 Regression

When $\mathbf{b}$ is not in the column space of $\mathbf{A}$, we can still find a vector $\mathbf{x}$ that comes the closest to solving $\mathbf{Ax} = \mathbf{b}$.

Figure 3 shows a simple example in which $\mathbf{A}$ and $\mathbf{b}$ are both 3x1 vectors and x is a scalar. In other words, $\mathbf{b}$ is a point in 3-space and the column space of $\mathbf{A}$ is a line in 3-space (different values of x put you at different points along that line). Everything that we do next is true for $\mathbf{A}$'s and $\mathbf{b}$'s of any size or length. But it's worthwhile keeping this simple picture in mind, because we can't draw diagrams in higher dimensions.

Here's the game. I give you $\mathbf{A}$ and $\mathbf{b}$. You pick an $\hat{\mathbf{x}}$ such that $\mathbf{A}\hat{\mathbf{x}}$ comes as close as possible to $\mathbf{b}$. That is, you pick $\hat{\mathbf{x}}$ to minimize:

$$||\mathbf{Ax} - \mathbf{b}||^2.$$

The solution, of course, is the orthogonal projection of $\mathbf{b}$ onto $\mathbf{Ax}$. This is clear for the simple example in figure 3; the shortest distance from the point, $\mathbf{b}$, to the line, $\mathbf{A}x$, is the perpendicular distance.

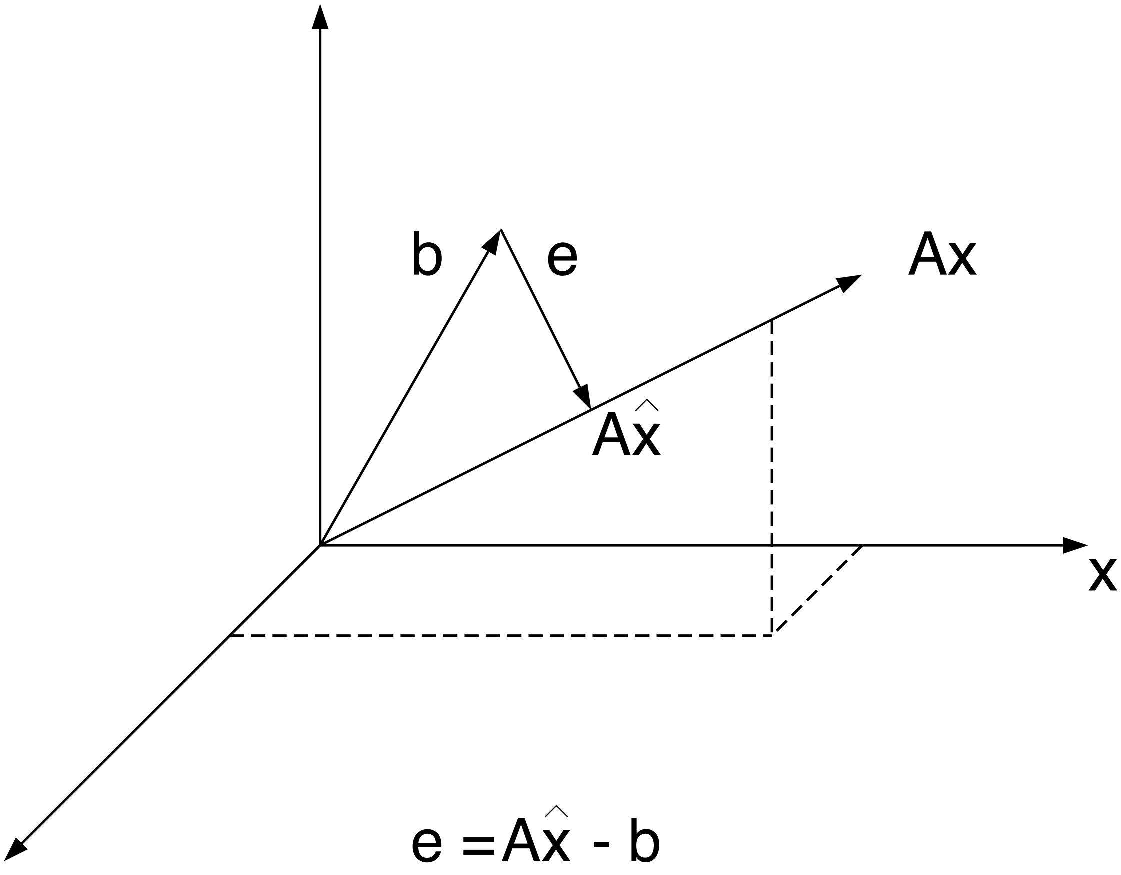

Figure 1: Regression: find the vector $\hat{\mathbf{x}}$ such that $\mathbf{A}\hat{\mathbf{x}}$ comes as close as possible to $\mathbf{b}$. The solution is the orthogonal projection of $\mathbf{b}$ onto $\mathbf{A}x$. The error vector $\mathbf{e}$ is perpendicular to $\mathbf{A}\hat{\mathbf{x}}$.

Given a value for $\hat{\mathbf{x}}$, the *error vector* (labeled in the figure) is given by the difference:

$$\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}.$$

Given the right choice of $\hat{\mathbf{x}}$, this error vector is perpendicular to the column space of $\mathbf{A}$:

$$(\mathbf{A}\mathbf{x})^t(\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}) = 0 \quad \text{or} \quad \mathbf{x}^t[\mathbf{A}^t\mathbf{A}\hat{\mathbf{x}} - \mathbf{A}^t\mathbf{b}] = 0.$$

This is true for every $\mathbf{x}$, i.e., it is true for all $\mathbf{y} = \mathbf{A}\mathbf{x}$ in the column space of $\mathbf{A}$. There is only one way in which this can happen: The vector in square brackets has to be the zero vector:

$$[\mathbf{A}^t\mathbf{A}\hat{\mathbf{x}} - \mathbf{A}^t\mathbf{b}] = \mathbf{0},$$

i.e.,

$$\hat{\mathbf{x}} = (\mathbf{A}^t\mathbf{A})^{-1}(\mathbf{A}^t\mathbf{b}).$$

Here the matrix $(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t$ is called the *pseudo-inverse* of $\mathbf{A}$.

We have to worry about one more thing: When is $\mathbf{A}^t\mathbf{A}$ invertible? One case is easy: If $\mathbf{A}$ has linearly independent columns, then $\mathbf{A}^t\mathbf{A}$ is a square, symmetric, and invertible matrix.

What if $\mathbf{A}$ is not full rank (i.e., what if it does not have linearly independent columns)? We still might be OK. Imagine that $\mathbf{A}$ is a 3x2 matrix, but that the two columns are identical. The columns are clearly not independent. We can proceed by dumping one (either one) of them. Make $\mathbf{A}'$, a new 3x1 vector that contains the first column of $\mathbf{A}$. Proceed as above to find the scalar $\hat{x}$ to minimize: $||\mathbf{A}'x - \mathbf{b}||^2$. The solution to our original problem is simply: $\hat{\mathbf{x}} = (\hat{x}, 0)^t$.

Generally, we use the SVD to compute the pseudo-inverse, $\mathbf{A}^\#$ of a matrix, $\mathbf{A}$. The SVD decomposes:

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^t,$$

where **S** is an NxM matrix with non-zero entries, s_i, along the main diagonal. The pseudo-inverse is computed:

$$\mathbf{A}^\# = \mathbf{V}\mathbf{S}^\#\mathbf{U},$$

where $\mathbf{S}^\#$ is an MxN matrix with non-zero entries, $s_i^\#$:

$$s_i^\# = \begin{cases} \frac{1}{s_i} & \text{for } s_i > t \\ 0 & \text{otherwise} \end{cases}$$

Here t is a threshold that is typically chosen based on the round-off error of the computer you're using. If one of the s_i's is very small (essentially zero), then the matrix is not full rank and we want to ignore one of the columns. Using the SVD is a very reliable method for inverting a matrix.

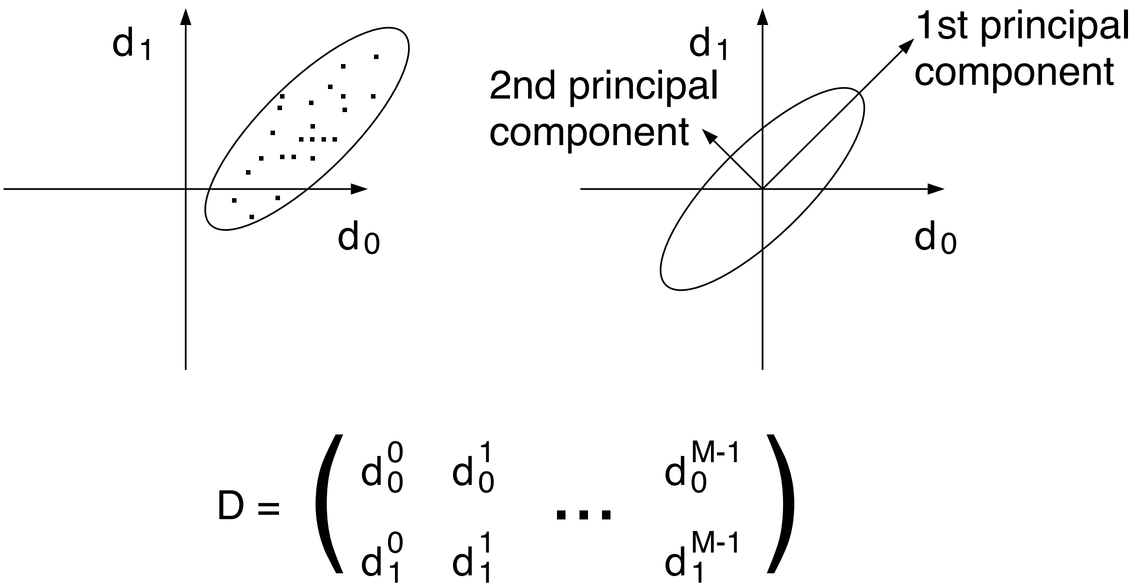

Figure 2: Top-left: Scatter plot of 2-dimensional data set. There are M data points. Each data point is represent by a vector, $\mathbf{d} = (d_0^i, d_1^i)$, for i from 1 to M. Top-right: Scatter plot of data set after subtracting the mean. The first principal component is a unit vector in the elongated direction of the scatter plot.

4 Covariance, Eigenvectors, and Principal Components Analysis

This section of the handout deals with multi-dimensional data sets, and explains principal component analysis (also called the Karhonen-Loeve Transform).

Each data point represents one test condition, e.g., a data point might be the height and weight of a person represented as a vector: (height,weight). Or each data point might include the intensity values of all of the pixels in an image represented as a very long vector $(p_1, p_2, \ldots, p_N)$, where p_i is the intensity at the ith pixel.

Let's say we have M such data points, each an N-dimensional vector. The average of all the data (the average of the columns) is itself an N-dimensional vector. Subtracting the mean vector from each data vector gives us a new set of N-dimensional vectors. In what follows, we'll be dealing exclusively with these new vectors (after subtracting out the mean); doing it this way makes the notation much simpler. We can put these vectors into a big NxM matrix, $\mathbf{D}$, where each column of this matrix corresponds to a single data point. Figure 4 shows an example in which N=2.

The covariance of the data is:

$$\mathrm{cov}(\mathbf{D}) = \tfrac{1}{M}\mathbf{D}\mathbf{D}^t.$$

This definition of the covariance is correct for data sets of arbitrary dimenion, N. For our simple 2-dimensional data set, we get:

$$\begin{aligned}\mathrm{cov}(\mathbf{D}) &= \tfrac{1}{M}\left(\begin{array}{ccc} d_0^0 & \cdots & d_0^{M-1} \\ d_1^0 & \cdots & d_1^{M-1}\end{array}\right)\left(\begin{array}{cc} d_0^0 & d_1^0 \\ \vdots & \vdots \\ d_0^{M-1} & d_1^{M-1}\end{array}\right) \\ &= \left(\begin{array}{cc} c_{0,0} & c_{0,1} \\ c_{1,0} & c_{1,1}\end{array}\right).\end{aligned}$$

Here $c_{0,0}$ is the variance of the data along the d_0 axis, $c_{1,1}$ is the variance of the data along the d_1 axis, and $c_{1,0} = c_{0,1}$ is the covariance (or correlation coefficient):

$$\begin{aligned} c_{0,0} &= \tfrac{1}{M}\sum_{i=1}^{M-1}\left(d_0^i\right)^2 \\ c_{1,1} &= \tfrac{1}{M}\sum_{i=1}^{M-1}\left(d_1^i\right)^2 \\ c_{0,1} = c_{1,0} &= \tfrac{1}{M}\sum_{i=1}^{M-1}\left(d_0^i d_1^i\right).\end{aligned}$$

For the data set illustrated in figure 4, the diagonal ($c_{0,0}$ and $c_{1,1}$) entries in the covariance matrix are about equal, and the off-diagonal($c_{0,1}$ and $c_{1,0}$) entries are reasonably large (about equal to 0.8) because the data is pretty elongated in that direction.

Using the SVD,

$$\mathbf{D} = \mathbf{U}\mathbf{S}\mathbf{V}^t,$$

the columns of $\mathbf{U}$ span the column space of $\mathbf{D}$. It turns out that these are also the eigenvectors of the covariance matrix, $\mathbf{D}\mathbf{D}^t$.

An eigenvector, $\mathbf{e}$, of a square matrix, $\mathbf{A}$, satisfies:

$$\mathbf{A}\mathbf{e} = \lambda\mathbf{e},$$

for some scalar λ, that is called an eigenvalue. In other words, the direction of $\mathbf{e}$ is unchanged by passing it through the matrix; only the length will change.

Let $\mathbf{U}_i$ be the ith column of $\mathbf{U}$. For $\mathbf{U}_i$ to be an eigenvector of $\mathbf{D}\mathbf{D}^t$, we need:

$$\mathbf{D}\mathbf{D}^t\mathbf{U}_i = \lambda\mathbf{U}_i.$$

Using the SVD,

$$\begin{aligned}\mathbf{D}\mathbf{D}^t &= \mathbf{U}\mathbf{S}\mathbf{V}^t\mathbf{V}\mathbf{S}\mathbf{U}^t \\ &= \mathbf{U}\mathbf{S}^2\mathbf{U}^t.\end{aligned}$$

Since $\mathbf{V}$ is orthonormal, $\mathbf{V}^t\mathbf{V} = I$ is an identity matrix. Then,

$$\mathbf{D}\mathbf{D}^t\mathbf{U}_i = \mathbf{U}\mathbf{S}^2\mathbf{U}^t\mathbf{U}_i$$

$$
\begin{aligned}
&= \mathbf{U}\mathbf{S}^2 \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \\
&= \mathbf{U} \begin{pmatrix} 0 & & & & \\ & \ddots & & & \\ & & s_i^2 & & \\ & & & \ddots & \\ & & & & 0 \end{pmatrix} \\
&= s_i^2 \mathbf{U}_i.
\end{aligned}
$$

The second line follows from the fact that $\mathbf{U}$ is orthonormal. The last line says that the square of the ith element of $\mathbf{S}$ is the eigenvalue corresponding to the ith eigenvector of $\mathbf{D}\mathbf{D}^t$, and that the eigenvector is given by the ith column of $\mathbf{U}$.

The eigenvector corresponding to the largest eigenvalue is called the first principal component of the covariance matrix. Most implementations of the SVD return the eigenvalues and eigenvectors in order so $\mathbf{U}_0$ (the first column of $\mathbf{U}$) is the first principal component. The second principal component is the eigenvector with the second largest eigenvalue (that is, the second column of $\mathbf{U}$), and so on.

The key property of the principal components (that we state here without proof) is that they capture the structure of the data. In particular, the first principal component is the unit vector with the largest projection onto the data set, i.e., the following expression is maximized for $\mathbf{e} = \mathbf{U}_0$:

$$||\mathbf{e}^t \mathbf{D}||^2.$$

As usual, it is helpful to draw the shape of the matrices on a piece of paper. For the simple example illustrated in figure 4, the first principal component is a unit vector in the (1,1) direction, the elongated direction of the scatter plot. In this case, $\mathbf{D}$ is a 2xM matrix, and $\mathbf{U}_0$ is 2x1 vector, so $\mathbf{U}_0^t\mathbf{D}$ is a 1xM row vector. Each element of this row vector is the length of the projection of a single data point onto the unit vector $\mathbf{U}_0$.

The point is that you know a lot about a data point by knowing only the projection of that data point onto $\mathbf{U}_0$. In fact, if you have to characterize each data point with only 1 number, you should use the projection onto the first principal component. That will give you the best approximation for the entire data set (given that you are only using 1 number for each data point).

What about the other principal components, the other columns of $\mathbf{U}$. The second principal component is the unit vector, in the subspace orthogonal to the first principal component, that has the largest projection onto the data set. The third principal component is the unit vector, in the subspace orthogonal to both the first and second principal components, that has the largest projection. And so on. If you have to characterize each data point with 3 numbers, you should use the projections onto each of the first three principal components.