L25SVD

December 9, 2015

1 The Singular Value Decomposition

Today we'll study the most useful decomposition in applied Linear Algebra.

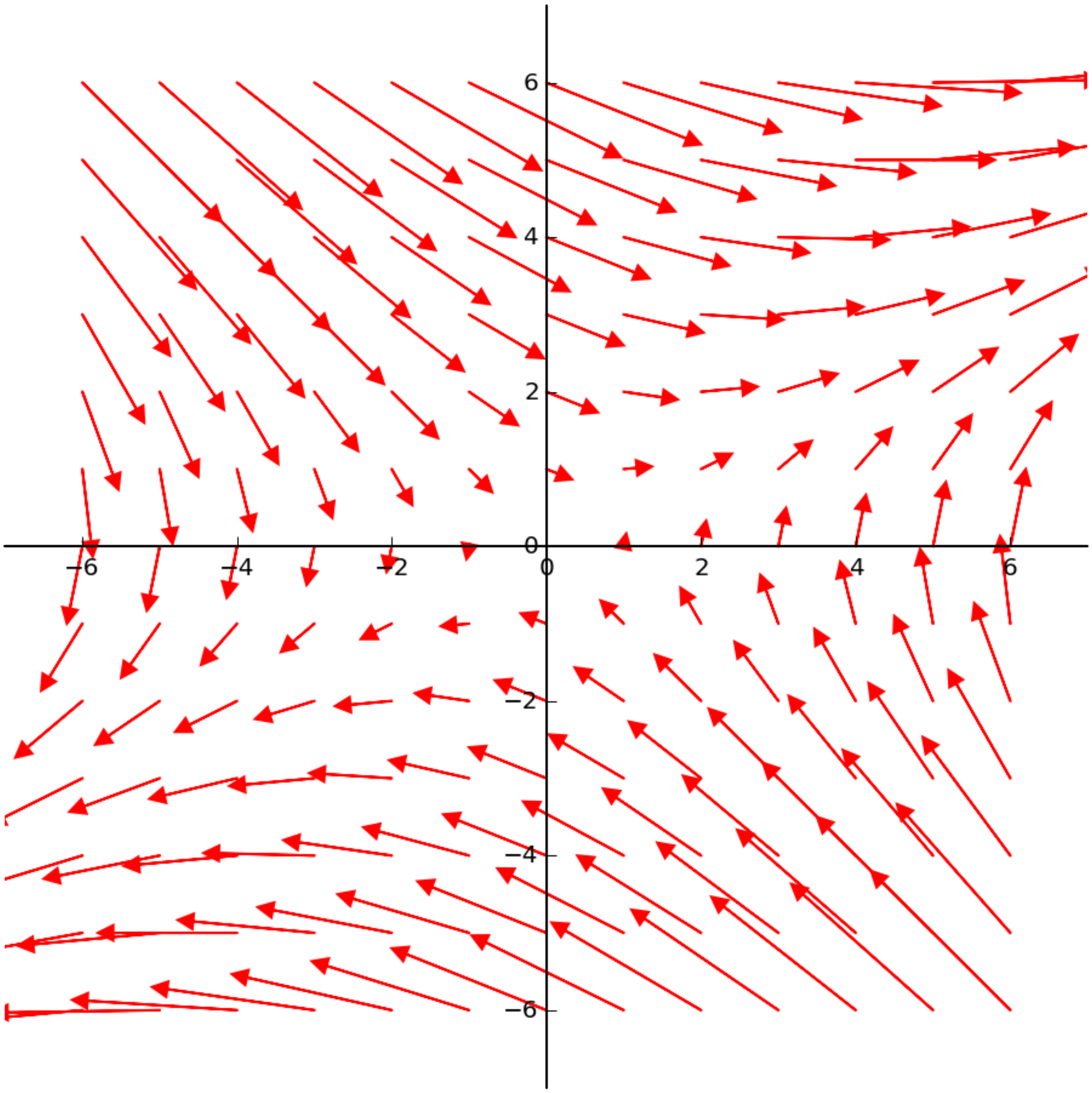
Pretty exciting, eh?

The singular value decomposition is a matrix factorization.

EVERY matrix has a singular value decomposition.

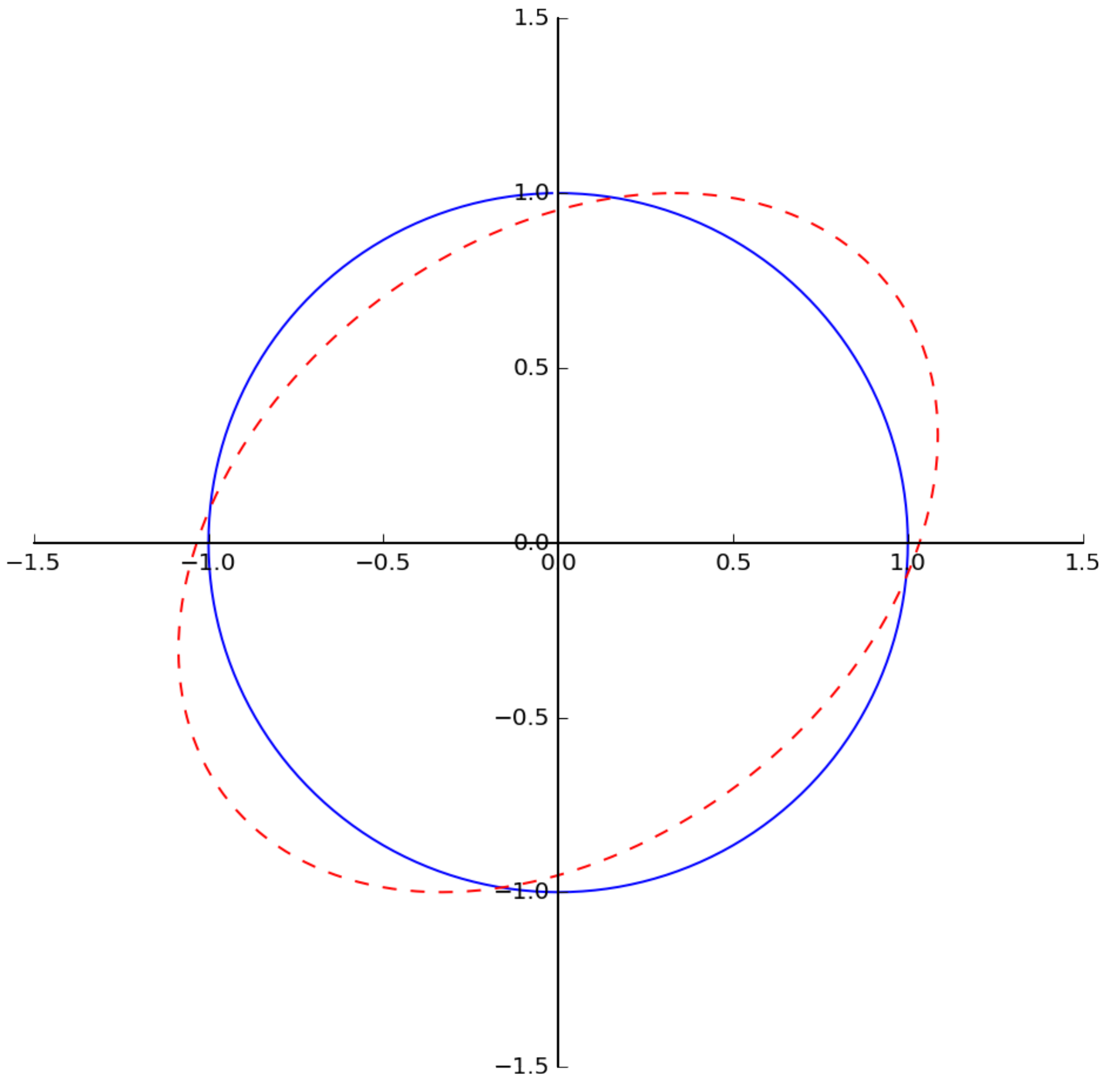
The singular value decomposition (let's just call it SVD) is based on a very simple idea, which is closely related to eigendecomposition.

Recall: the eigenvalues of a matrix A measure the amount that A "stretches or shrinks" certain special vectors (the eigenvectors).



For example, if $A\mathbf{x} = \lambda\mathbf{x}$ and $\|x\| = 1$, then

$$\|A\mathbf{x}\| = \|\lambda\mathbf{x}\| = |\lambda|\,\|\mathbf{x}\| = |\lambda|.$$



If λ_1 is the eigenvalue with the greatest magnitude, then a corresponding unit eigenvector $\mathbf{v}_1$ identifies a direction in which the stretching effect of A is greatest.

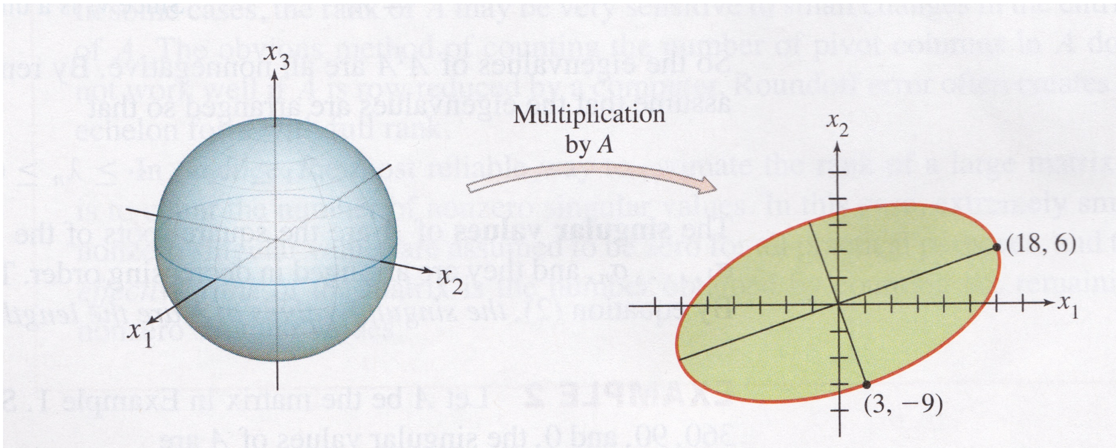
That is, over all unit vectors $\mathbf{x}$, the length of $A\mathbf{x}$ is maximized when $\mathbf{x} = \mathbf{v}_1$.

In which case, $\|A\mathbf{v}_1\| = |\lambda_1|$.

Now let's see by example how we can extend this idea to **arbitrary** (non-square) matrices.

Example.

If $A = \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix}$, then the linear transformation $\mathbf{x} \mapsto A\mathbf{x}$ maps the unit sphere $\{\mathbf{x} : \|\mathbf{x}\| = 1\}$ in $\mathbb{R}^3$ onto an ellipse in $\mathbb{R}^2$, as shown here:



Problem. Find the unit vector $\mathbf{x}$ at which the length $\|A\mathbf{x}\|$ is maximized, and compute this maximum length.

Solution.

The quantity $\|A\mathbf{x}\|^2$ is maximized at the same $\mathbf{x}$ that maximizes $\|A\mathbf{x}\|$, and $\|A\mathbf{x}\|^2$ is easier to study.

Observe that

$$\|A\mathbf{x}\|^2 = (A\mathbf{x})^T(A\mathbf{x})$$

$$= \mathbf{x}^T A^T A\mathbf{x}$$

$$= \mathbf{x}^T (A^T A)\mathbf{x}$$

Now, $A^T A$ is a symmetric matrix.

So the above is a quadratic form, and we are seeking to maximize it subject to the constraint $\|x\| = 1$.

As we learned in the last lecture, the maximum value subject to the constraint is the largest eigenvalue λ_1 of $A^T A$.

Also, the maximum is attained at a unit eigenvector of $A^T A$ corresponding to λ_1.

For the matrix A in the example,

$$A^T A = \begin{bmatrix} 4 & 8 \\ 11 & 7 \\ 14 & -2 \end{bmatrix} \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix} = \begin{bmatrix} 80 & 100 & 40 \\ 100 & 170 & 140 \\ 40 & 140 & 200 \end{bmatrix}.$$

The eigenvalues of $A^T A$ are $\lambda_1 = 360, \lambda_2 = 90$, and $\lambda_3 = 0$.

The corresponding unit eigenvectors are, respectively,

$$\mathbf{v}_1 = \begin{bmatrix} 1/3 \\ 2/3 \\ 2/3 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} -2/3 \\ -1/3 \\ 2/3 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 2/3 \\ -2/3 \\ 1/3 \end{bmatrix}.$$

For $\|\mathbf{x}\| = 1$, the maximum value of $\|A\mathbf{x}\|$ is $\|A\mathbf{v}_1\| = \sqrt{360}$.

This example shows that the key to understanding the effect of A on the unit sphere in $\mathbb{R}^3$ is to examime the quadratic form $\mathbf{x}^T(A^T A)\mathbf{x}$.

In fact, the entire geometric behavior of the transformation $\mathbf{x} \mapsto A\mathbf{x}$ is captured by this quadratic form.

1.1 The Singular Values of a Matrix

Let A be an arbtrary $m \times n$ matrix.

Notice that $A^T A$ is symmetric. So, it can be orthogonally diagonalized (as we saw in the last lecture).

So let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of $A^T A$, and let $\lambda_1, \ldots, \lambda_n$ be the corresponding eigenvalues of $A^T A$.

Then, for any eigenvector $\mathbf{v}_i$,

$$\|A\mathbf{v}_i\|^2 = (A\mathbf{v}_i)^T A\mathbf{v}_i = \mathbf{v}_i^T A^T A\mathbf{v}_i$$

$$= \mathbf{v}_i^T(\lambda_i)\mathbf{v}_i$$

(since $\mathbf{v}_i$ is an eigenvector of $A^T A$)

$$= \lambda_i$$

(since $\mathbf{v}_i$ is a unit vector.)

So the eigenvalues of $A^T A$ are all nonnegative

(that is: $A^T A$ is positive semidefinite).

We can therefore renumber the eigenvalues so that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0.$$

Definition. The **singular values** of A are the square roots of the eigenvalues of $A^T A$. They are denoted by $\sigma_1, \ldots, \sigma_n$, and they are arranged in decreasing order.

That is, $\sigma_i = \sqrt{\lambda_i}$ for $i = 1, \ldots, n$.

By the above argument, **the singular values of A are the lengths of the vectors** $A\mathbf{v}_1, \ldots, A\mathbf{v}_n$.

Now: we know that vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are an orthogonal set because they are eigenvectors of the symmetric matrix $A^T A$.

However, it's **also** the case that $A\mathbf{v}_1, \ldots, A\mathbf{v}_n$ are an orthogonal set.

... a fact which is key to the SVD.

Let's prove it.

Theorem. Suppose $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of $A^T A$, arranged so that the corresponding eigenvalues of $A^T A$ satisfy $\lambda_1 \geq \cdots \geq \lambda_n$, and suppose A has r nonzero singular values. Then $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_r\}$ is an orthogonal basis for Col A, and rank $A = r$.

Proof. Recall that what we need to do is establish that $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_r\}$ is a (orthogonal) linearly independent set whose span is Col A.

Because $\mathbf{v}_i$ and $\mathbf{v}_j$ are orthogonal for $i \neq j$,

$$(A\mathbf{v}_i)^T(A\mathbf{v}_j) = \mathbf{v}_i^T A^T A\mathbf{v}_j = \mathbf{v}_i^T(\lambda_j \mathbf{v}_j) = 0.$$

So $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_n\}$ is an orthogonal set.

Furthermore, since the lengths of the vectors $A\mathbf{v}_1, \ldots, A\mathbf{v}_n$ are the singular values of A, and since there are r nonzero singular values, $A\mathbf{v}_i \neq \mathbf{0}$ if and only if $1 \leq i \leq r$.

So $A\mathbf{v}_1, \ldots, A\mathbf{v}_r$ are a linearly independent set (because they are orthogonal and all nonzero), and clearly they are each in Col A.

Finally, we just need to show that Span $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_r\}$ = Col A.

To do this we'll show that for any $\mathbf{y}$ in Col A, we can write $\mathbf{y}$ in terms of $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_r\}$:

Say $\mathbf{y} = A\mathbf{x}$.

Because $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis for $\mathbb{R}^n$, we can write $\mathbf{x} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$, so

$$\mathbf{y} = A\mathbf{x} = c_1 A\mathbf{v}_1 + \cdots + c_r A\mathbf{v}_r + \cdots + c_n A\mathbf{v}_n.$$

$$= c_1 A\mathbf{v}_1 + \cdots + c_r A\mathbf{v}_r.$$

(because $A\mathbf{v}_i = \mathbf{0}$ for $i > r$).

In summary: $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_n\}$ is an (orthogonal) linearly independent set whose span is Col A, so it is an (orthogonal) basis for Col A.

Notice that we have also proved that rank $A = \dim \text{Col } A = r$.

In other words, if A has r nonzero singular values, A has rank r.

1.2 The Singular Value Decomposition

Note that the domain of $A\mathbf{x}$ is $\mathbb{R}^n$ and the range of $A\mathbf{x}$ is Col A.

So what we have proved is that the eigenvectors of A^TA are rather special.

We have proved that the set $\{\mathbf{v}_i\}$ is an orthogonal basis for the domain of $A\mathbf{x}$, and $\{A\mathbf{v}_i\}$ is an orthogonal basis for the range of $A\mathbf{x}$.

Now we can define the SVD.

Theorem. Let A be an $m \times n$ matrix with rank r. Then there exists an $r \times r$ matrix Σ whose diagonal entries are the r nonzero singular values of A, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, and there exist an $m \times r$ orthogonal matrix U and an $n \times r$ orthogonal matrix V such that

$$A = U\Sigma V^T.$$

Any factorization $A = U\Sigma V^T$, with U and V orthogonal and Σ a diagonal matrix is called a **singular value decomposition (SVD)** of A.

The columns of U are called the **left singular vectors** and the columns of V are called the **right singular vectors** of A.

We have built up enough tools now that the proof is quite straightforward.

Proof. Let λ_i and $\mathbf{v}_i$ be the eigenvalues and eigenvectors of A^TA, and $\sigma_i = \sqrt{\lambda_i}$.

As we have seen, $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_r\}$ is an orthogonal basis for Col A.

Normalize each $A\mathbf{v}_i$ to obtain an orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$, where

$$\mathbf{u}_i = \frac{1}{\|A\mathbf{v}_i\|} = \frac{1}{\sigma_i}A\mathbf{v}_i$$

Then

$$A\mathbf{v}_i = \sigma_i\mathbf{u}_i \quad (1 \leq i \leq r)$$

So

$$AV = \begin{bmatrix} A\mathbf{v}_1 & \cdots & A\mathbf{v}_r \end{bmatrix} = \begin{bmatrix} \sigma_1\mathbf{u}_1 & \cdots & \sigma_r\mathbf{u}_r \end{bmatrix} = U\Sigma.$$

Now, V is an orthogonal matrix, so

$$U\Sigma V^T = AVV^T = A.$$

1.3 Approximating a Matrix

One way to think of the SVD is that it gives tools for approximating a matrix with another matrix.

To talk about when one matrix "approximates" another, we need a "length" for matrices.

We will use the **Frobenius norm** which is just the usual norm, treating the matrix as if it were a vector.

In other words, the definition of the Frobenius norm of A, denoted $\|A\|_F$, is:

$$\|A\|_F = \sqrt{\sum a_{ij}^2}.$$

The approximations we'll discuss are **low-rank** approximations.

Recall that the rank of a matrix A is the largest number of linearly independent columns of A.

Let's define the **rank-k approximation** to A:

When $k <$ rank A, the rank-k approximation to A is the closest rank-k matrix to A, i.e.,

$$A^{(k)} = \arg \min_{\text{rank } B=k} \|A - B\|_F.$$

Note that this matrix may take up **much** less space than the orignal A.

$$
m\left\{ \vphantom{\begin{bmatrix} \vdots \\ \vdots \\ \mathbf{a_1} \\ \vdots \\ \vdots \end{bmatrix}} \right.
\overbrace{\begin{bmatrix} \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \mathbf{a_1} & \mathbf{a_2} & \dots & \mathbf{a_n} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \end{bmatrix}}^{n}
=
\overbrace{\begin{bmatrix} \vdots & \vdots \\ \vdots & \vdots \\ \sigma_1\mathbf{u}_1 & \sigma_k\mathbf{u}_k \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix}}^{k}
\times
\begin{bmatrix} \dots & \dots & \mathbf{v}_1 & \dots & \dots \\ \dots & \dots & \mathbf{v}_k & \dots & \dots \end{bmatrix}
$$

The rank-k approximation takes up space $(m+n)k$ while A itself takes space mn.

For example, if $k = 10$ and $m = n = 1000$, then the rank-k approximation takes space $20000/1000000 =$ 2% of A.

Here is (one of many) remarkable facts about the SVD:

The best rank-k approximation to any matrix can be found via the SVD.

In fact, for an $m \times n$ matrix A, the SVD does two things:

1. It gives the best rank-k approximation to A for **every** k up to the rank of A.
2. It gives the **distance** of the best rank-k approximation $A^{(k)}$ from A for each k.

In terms of the singular value decomposition,

1) The best rank-k approximation to A is formed by taking

   - \$U' = \$ the k leftmost columns of U,
   - \$$\Sigma' =$ \$*the*k$\times k$ upper left submatrix of Σ, and
   - \$V'= \$ the k leftmost columns of V, and constructing

$$A^{(k)} = U'\Sigma'(V')^T.$$

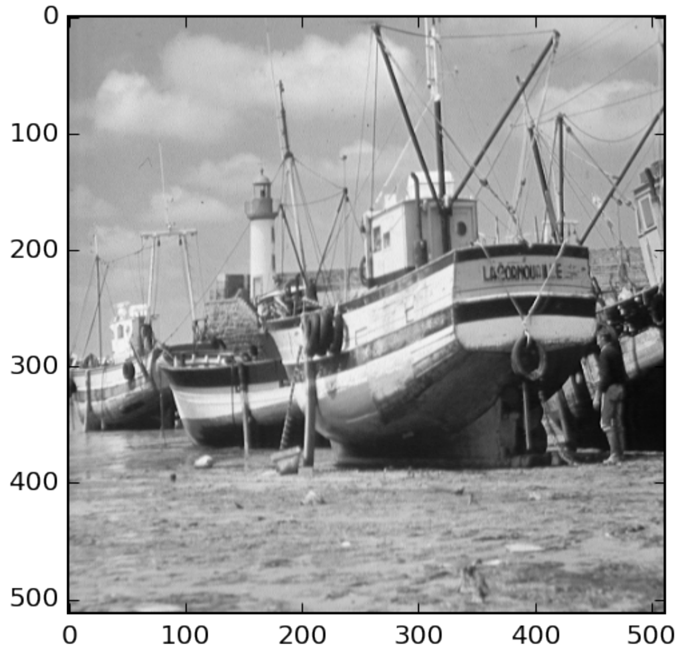
2) The distance (in Frobenius norm) of the best rank-k approximation $A^{(k)}$ from A is equal to $\sqrt{\sum_{i=k+1}^{r} \sigma_i^2}$.

What this means is that if, beyond some k, all of the singular values are small, then A can be closely approximated by a rank-k matrix.

Example: signal compression.

Image data is often **approximately low-rank.**

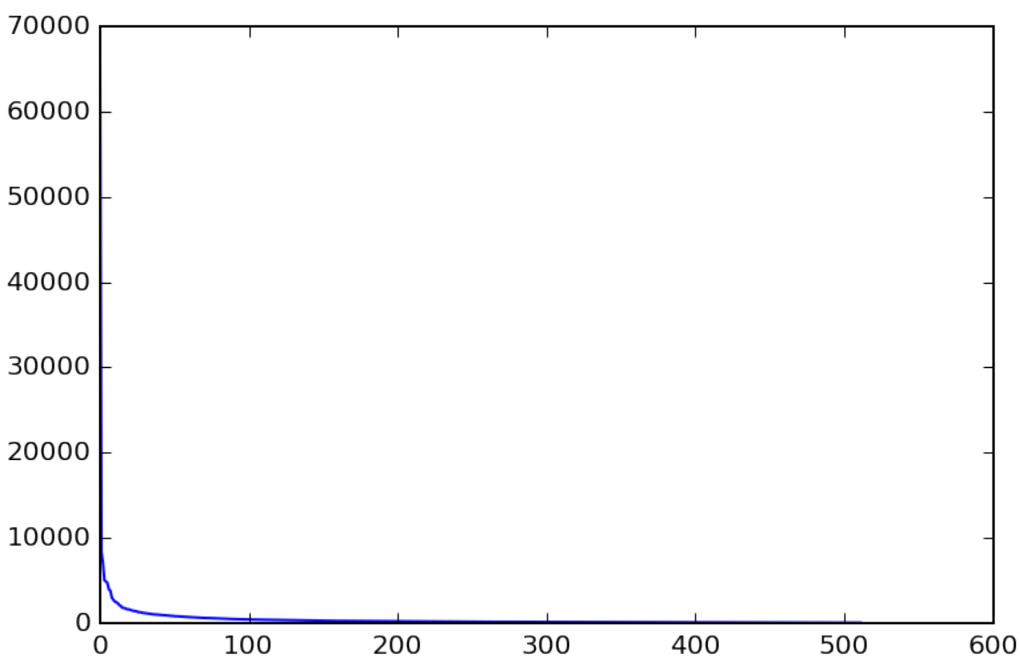
For example, here is a photo, which is really a 512×512 matrix:

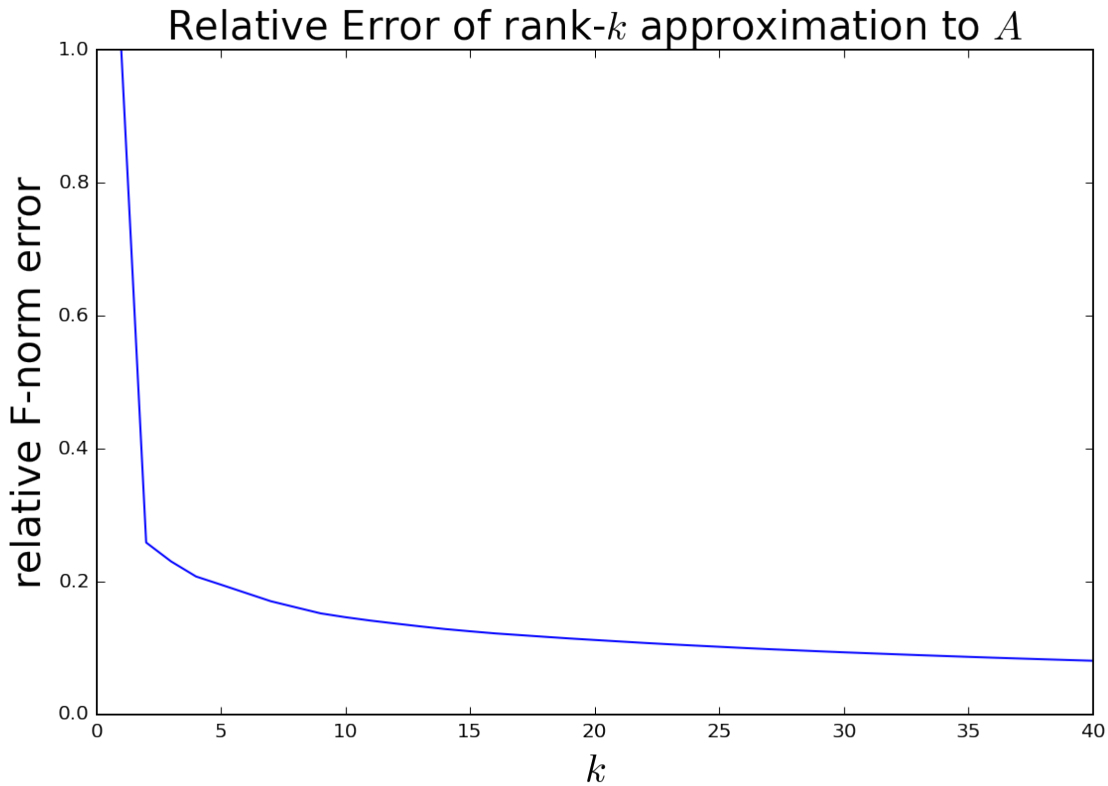


Let's look at its singular values (often called the matrix's "spectrum"):

```
In [34]: u,s,vt=np.linalg.svd(boat,full_matrices=False)
         plt.figure()
         plt.plot(s)
```

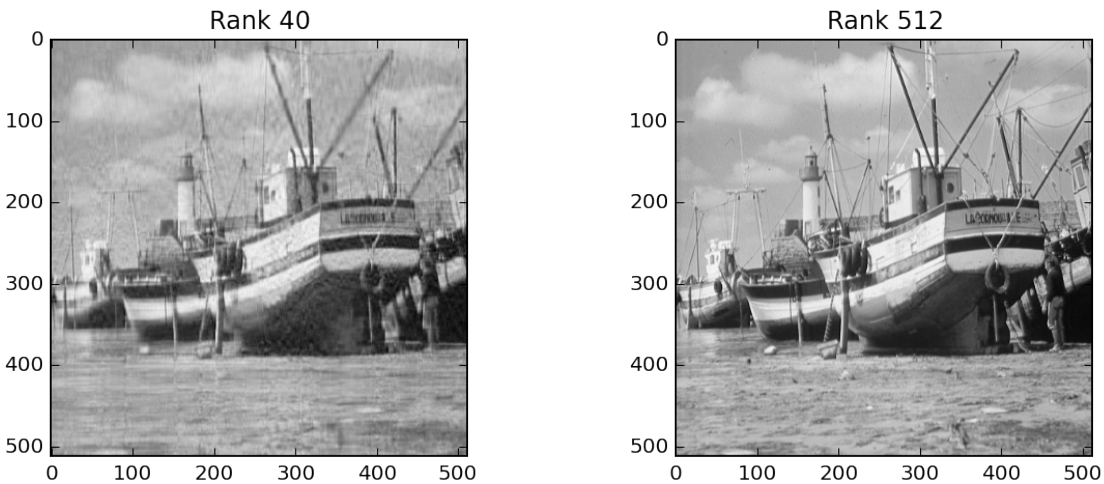
```
Out[34]: [<matplotlib.lines.Line2D at 0x113d4ddd8>]
```





This matrix has rank of 512. But its "effective" rank is low, perhaps 40.
Let's find the closest rank-40 matrix and view it.

In [45]:
```
# construct a rank-n version of the boat
scopy = s.copy()
rank = 40
scopy[rank:]=0
boatApprox = u.dot(np.diag(scopy)).dot(vt)
#
plt.figure(figsize=(9,6))
plt.subplot(1,2,1)
plt.imshow(boatApprox,cmap = cm.Greys_r)
plt.title('Rank {}'.format(rank))
plt.subplot(1,2,2)
plt.imshow(boat,cmap = cm.Greys_r)
plt.title('Rank 512')
plt.subplots_adjust(wspace=0.5)
print('')
```



Note that the rank-40 boat takes up only $40/512 =$ **8% of the space as the original image!**

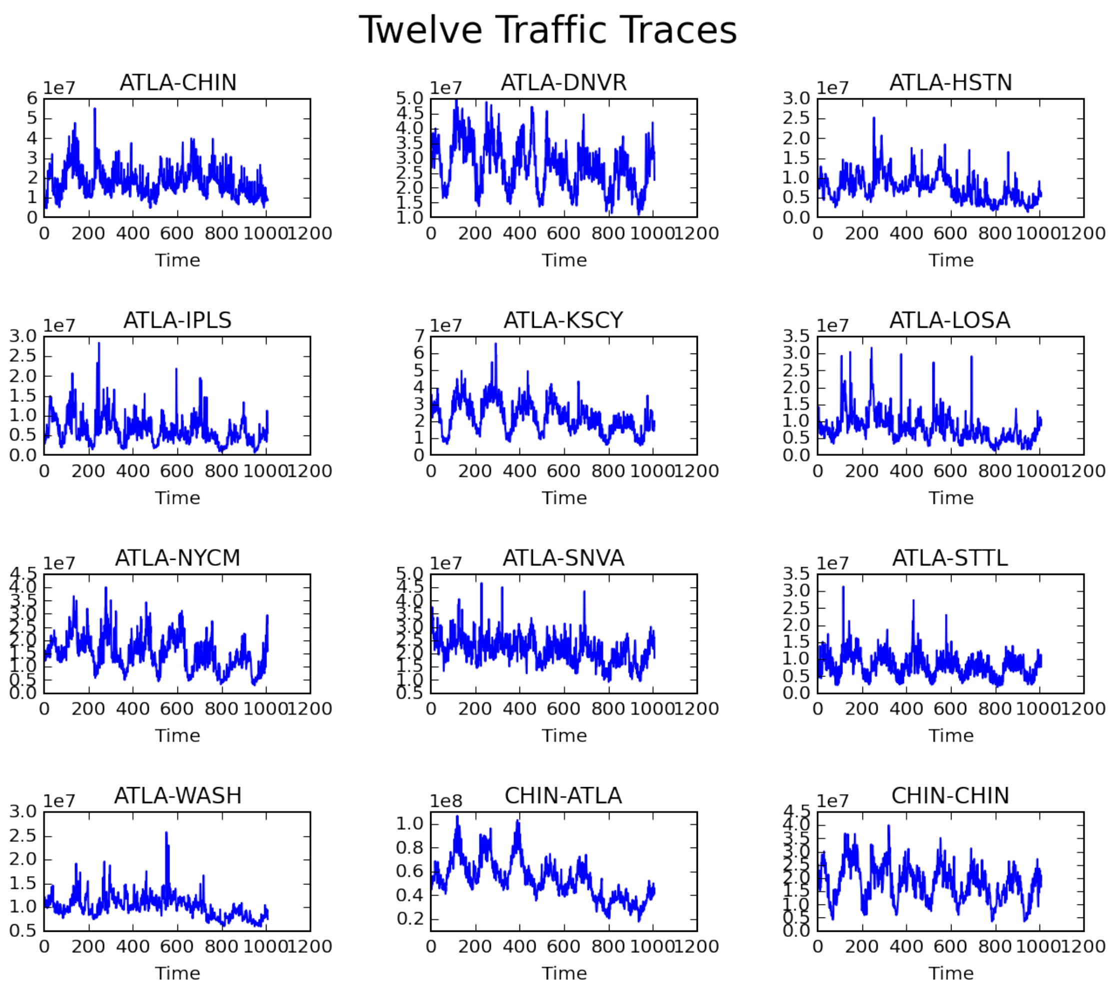
This general principle is what makes image, video, and sound compression effective.

When you watch HDTV, or listen to an MP3, or look at a JPEG image, these signals have been compressed using the fact that they are basically **low-rank** matrices.

Example: Pattern extraction.

Another remarkable feature of the SVD is that it **automatically extracts common patterns** from a set of data.

Here is an example: data traffic flowing over a network.



Each traffic trace is a column of A.

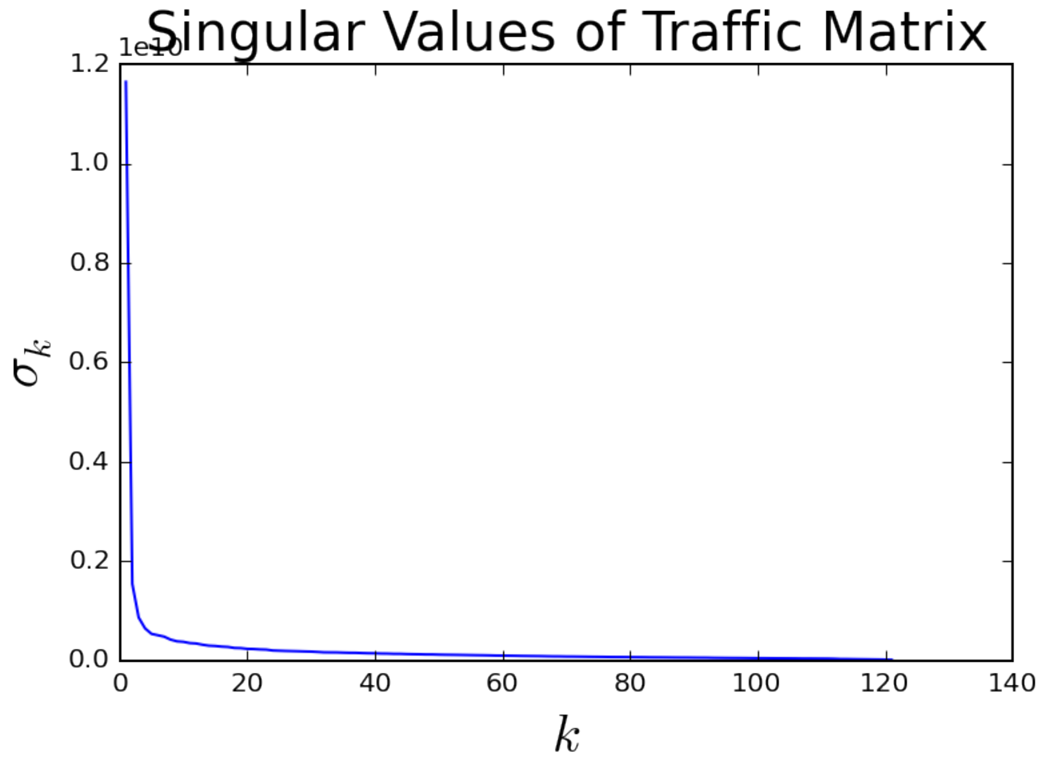
$$A \approx U'\Sigma'(V')^T$$

In this interpretation, we think of each column of A as a combination of the columns of U'.
Let's use as our example $\mathbf{a}_1$, the first column of A
The equation above tells us that

$$\mathbf{a}_1 \approx v_{11}\sigma_1\mathbf{u}_1 + v_{12}\sigma_2\mathbf{u}_2 + \cdots + v_{1k}\sigma_k\mathbf{u}_k.$$

In other words, $\mathbf{u}_1$ (the first column of U) is the "strongest" pattern occurring in A, and its strength is measured by σ_1.

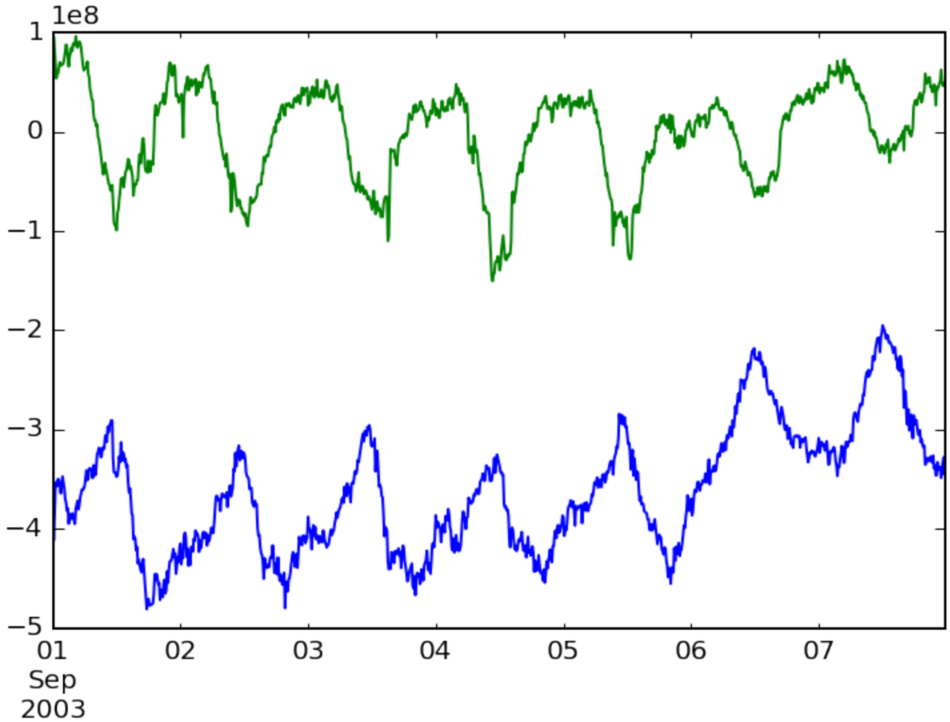
In [53]:
```
u,s,vt = np.linalg.svd(Atraf)
fig = plt.figure(figsize=(6,4))
plt.plot(range(1,1+len(s)),s)
plt.xlabel(r'$k$',size=20)
plt.ylabel(r'$\sigma_k$',size=20)
plt.title(r'Singular Values of Traffic Matrix',size=20)
print('')
```



Here is an view of the first two columns of $U\Sigma$ for the traffic matrix data:

```
In [11]: u,s,vt = np.linalg.svd(Atraf,full_matrices=False)
         uframe = pd.DataFrame(u.dot(np.diag(s)),index=pd.date_range('9/1/2003',freq='10min',periods=10
         uframe[0].plot()
         uframe[1].plot()

Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x110414f60>
```



1e8
1
0
−1
−2
−3
−4
−5
01
Sep
2003
02
03
04
05
06
07