10-Low-Rank-and-SVD

October 12, 2016

1 Low Rank Approximation and the SVD

Today, we move on from clustering.

However, let's look back and try to put clustering into a larger context.

1.1 Models are simplifications

One way of thinking about clustering is that we are building a **simplification** of the data.

That is, a model of the data that is simpler than the data.

In particular, instead of thinking of the data as thousands or millions of individual data points, we think of it in terms of a small number of clusters.

From this simpler description, we hope to gain **insight.**

For example, we hope to learn how restaurants in Las Vegas happen to locate near other restaurants of similar type.

There is an interesting question here: **why** does this process often lead to insight?

That is, why does it happen so often that a large dataset can be described in terms of a much simpler model?

I don't know.

By self-created (Moscarlop) - Own work, CC BY-SA 3.0, Link

However, I think that William of Ockham (c. 1300 AD) was on the right track.

He said:

> Non sunt multiplicanda entia sine necessitate

or, in other words:

> Entities must not be multiplied beyond necessity.

by which he meant:

> Among competing hypotheses, the one with the fewest assumptions should be selected.

Which has come to be known as "Occam's razor."

William was saying that it is more common for a set of observations to be determined by a simple process than a complex process.

In other words, the world is full of simple (but often hidden) patterns.

From which one can justify the observation that "clustering works suprisingly often."

1.2 Matrix Rank

Now we'll consider a (seemingly) very different approximation of data, applicable to data when it is in matrix form.

$$m \text{ data objects} \left\{ \vphantom{\begin{matrix} a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{m1} \end{matrix}} \right. \overbrace{\begin{bmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & \dots & a_{ij} & \dots & a_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mj} & \dots & a_{mn} \end{bmatrix}}^{n \text{ features} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \end{bmatrix}}$$

Data Type

Rows

Columns

Elements

Network Traffic

Sources

Destinations

Number of Bytes

Social Media

Users

time bins

Number of Posts/Tweets/Likes

Web Browsing

Users

Content Categories

Visit Counts/Bytes Downloaded

Web Browsing

Users

time bins

Visit Counts/Bytes Downloaded

Let's briefly review some linear algebra.

We'll consider an $m \times n$ real matrix A.

The **rank** of A is the **dimension of its column space.**

The dimension of a space is the smallest number of (linearly independent) vectors needed to span the space.

So the dimension of the column space of A is the smallest number of vectors that suffice to construct the columns of A.

Then the rank of A is the size of the smallest set $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p\}$ such that every $\mathbf{a}_i$ can be expressed as:

$$\mathbf{a}_i = c_{i1}\mathbf{u}_1 + c_{i2}\mathbf{u}_2 + \cdots + c_{ip}\mathbf{u}_p \quad i = 1, \dots, n.$$

The largest value that a matrix rank can take is $\min(m, n)$.

However it can happen that the rank of a matrix is **less** than $\min(m, n)$.

Now to store a matrix $A \in \mathbb{R}^{m \times n}$ we need to store mn values.

However, if A has rank k, it can be factored as $A = UV$, where $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{k \times n}$.

This only requires $k(m+n)$ values, which could be much smaller than mn.

$$n \text{ features} = \overbrace{\begin{bmatrix} \vdots & \vdots \\ \vdots & \vdots \\ \mathbf{u}_1 & \mathbf{u}_k \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix}}^{k} \times \begin{bmatrix} \cdots & \cdots & \mathbf{v}_1 & \cdots & \cdots \\ \cdots & \cdots & \mathbf{v}_k & \cdots & \cdots \end{bmatrix}$$

$$m \text{ data objects} \left\{ \vphantom{\begin{bmatrix} a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{m1} \end{bmatrix}} \right. \qquad \overbrace{\begin{bmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix}}$$

1.3 Low Effective Rank

In working with real data, what we generally want to do is **approximate** our data matrix A with a low-rank matrix $A^{(k)}$.

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

We will use the **Frobenius norm** which is just the usual ℓ_2 norm, treating the matrix as a vector.

The definition of the Frobenius norm of A, denoted $\|A\|_F$, is:

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

To quantify when one matrix is "close" to another, we use distance in Euclidean space:

$$\text{dist}(A, B) = \|A - B\|_F.$$

(where the Euclidean space is the mn-dimensional space of $m \times n$ matrices.)

Now we can define the **rank-k approximation** to A:

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

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

This can also be considered the best rank-k approximation to A in a least-squares sense.

Let's say we have $A^{(k)}$, a rank-k approximation to A.

By definition, there is a set U consisting of k vectors such that each column of $A^{(k)}$ can be expressed as a linear combination of vectors in U.

Let us abuse notation and also call the matrix formed by those vectors U.

So

$$A^{(k)} = UV^T$$

for some set of coefficients V^T that describe the linear combinations of U that yield the columns of $A^{(k)}$.

So U is $m \times k$ and V is $n \times k$.
If we approximate A by $A^{(k)}$, then the error we incur is:

$$\|A - A^{(k)}\|_F.$$

Hence, a rank-k approximation $A^{(k)}$ is valuable if

- $\|A - A^{(k)}\|_F$ is small compared to $\|A\|_F$, and
- k is small compared to m and n.

In that case we have achieved a simplification of the data without a great loss in accuracy.

1.4 Finding rank-k approximations

There is a celebrated method for finding the best rank-k approximation to any matrix: the **Singular Value Decomposition (SVD).**

> SVD is "the Rolls-Royce and the Swiss Army Knife of Numerical Linear Algebra."

Dianne O'Leary, MMDS '06
The singular value decomposition of a rank-r matrix A has the form:

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

where

1. U is $m \times r$
2. The columns of U are mutually orthogonal and unit length, ie., $U^TU = I$.
3. V is $n \times r$.
4. The columns of V are mutually orthogonal and unit length, ie., $V^TV = I$.
5. The matrix Σ is a $r \times r$ diagonal matrix, whose diagonal values are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

$$\begin{bmatrix} \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \mathbf{a_1} & \mathbf{a_2} & \dots & \mathbf{a_n} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \end{bmatrix} = \overbrace{\begin{bmatrix} \vdots & \vdots \\ \vdots & \vdots \\ \mathbf{u}_1 & \mathbf{u}_r \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix}}^{r} \times \begin{bmatrix} \sigma_1 & \\ & \sigma_r \end{bmatrix} \times \begin{bmatrix} \dots & \dots & \mathbf{v}_1 & \dots & \dots \\ \dots & \dots & \mathbf{v}_r & \dots & \dots \end{bmatrix}$$

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 approximation $A^{(k)}$ from A for each k.

In terms of the singular value decomposition,
The best rank-k approximation to A is formed by taking

- \$U' = \$ the k leftmost columns of U,

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

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

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

That is, if you construct $A^{(k)}$ as shown above, then:

$$\|A - A^{(k)}\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2$$

1.5 Low Effective Rank

Almost any data matrix A that one encounters will usually be **full rank**,

meaning that $\text{rank}\ A = min(m, n)$.

However, it is often the case that data matrices have **low effective rank.**

By this we mean that one can usefully approximate A by some $A^{(k)}$ for which $k \ll min(m, n)$.

For any data matrix, we can judge when this is the case by looking at its singular values, because the singular values tell us the distance to the nearest rank-k matrix.

1.6 Empirical Evidence

Let's see how this theory can be used in practice, and investigate some real data.

We'll look at OD flow traffic on the Abilene network:

Source: Internet2, circa 2005

```
In [65]: with open('data/net-traffic/AbileneFlows/odnames','r') as f:
             odnames = [line.strip() for line in f]
         dates = pd.date_range('9/1/2003',freq='10min',periods=1008)
         Atraf = pd.read_table('data/net-traffic/AbileneFlows/X',sep='  ',header=No
         Atraf.index = dates
         Atraf
```

```
Out[65]:                        ATLA-ATLA   ATLA-CHIN   ATLA-DNVR   ATLA-HSTN  \
         2003-09-01 00:00:00   8466132.0  29346537.0  15792104.0   3646187.0
         2003-09-01 00:10:00  20524567.0  28726106.0   8030109.0   4175817.0
         2003-09-01 00:20:00  12864863.0  27630217.0   7417228.0   5337471.0
         2003-09-01 00:30:00  10856263.0  32243146.0   7136130.0   3695059.0
         2003-09-01 00:40:00  10068533.0  30164311.0   8061482.0   2922271.0
         2003-09-01 00:50:00   6886434.0  26797124.0   9011267.0   3084852.0
         2003-09-01 01:00:00   4898240.0  30156627.0   8804811.0   3488872.0
         2003-09-01 01:10:00   6248053.0  29814965.0   9445386.0   4028627.0
         2003-09-01 01:20:00   7180983.0  30857813.0   8848281.0   4365325.0
         2003-09-01 01:30:00   7503555.0  31675685.0   8252067.0   4028992.0
         2003-09-01 01:40:00   8202708.0  31505085.0   7308166.0   4042905.0
         2003-09-01 01:50:00   5297330.0  38475090.0   8567962.0   5934312.0
```

```
2003-09-01 02:00:00   4627707.0  33481419.0  10043297.0   5943747.0
2003-09-01 02:10:00  13819493.0  36646755.0  11047352.0   7787379.0
2003-09-01 02:20:00  10886153.0  36284772.0  11072503.0   5386058.0
2003-09-01 02:30:00   8800323.0  27608102.0  12909895.0   4547743.0
2003-09-01 02:40:00   7354961.0  26580891.0  12901277.0   5329349.0
2003-09-01 02:50:00  11360523.0  26829203.0  11266330.0   6831101.0
2003-09-01 03:00:00  23585641.0  29074917.0  11750908.0   6196441.0
2003-09-01 03:10:00  23009442.0  30082310.0   9178331.0   6689856.0
2003-09-01 03:20:00  21235933.0  30749983.0   7953561.0   5898858.0
2003-09-01 03:30:00  19931353.0  37108075.0  11058469.0   4455230.0
2003-09-01 03:40:00  23376575.0  39977856.0  10484929.0   5250466.0
2003-09-01 03:50:00  19111330.0  34502235.0   9336522.0   7195325.0
2003-09-01 04:00:00  21357590.0  34776494.0   6885092.0  10779308.0
2003-09-01 04:10:00  27419003.0  36718681.0   4818045.0  12617814.0
2003-09-01 04:20:00  20081201.0  31682394.0   6052136.0  11728791.0
2003-09-01 04:30:00  21684723.0  32614579.0   6071353.0  12956449.0
2003-09-01 04:40:00  26603133.0  31575873.0   5365878.0  14954653.0
2003-09-01 04:50:00  23543383.0  33028816.0   4390253.0  12976114.0
...                         ...         ...         ...         ...
2003-09-07 19:00:00  19489715.0  31242150.0   3906168.0   4775617.0
2003-09-07 19:10:00  20457357.0  36109567.0   5205077.0   4340259.0
2003-09-07 19:20:00   8295874.0  38223524.0   5354467.0   4205196.0
2003-09-07 19:30:00   8664180.0  38545654.0   5233399.0   4108983.0
2003-09-07 19:40:00   8825123.0  37247038.0   3982497.0   5167667.0
2003-09-07 19:50:00   7538140.0  36333668.0   3199209.0   4594949.0
2003-09-07 20:00:00   7738575.0  35547964.0   3981707.0   4970804.0
2003-09-07 20:10:00  16303853.0  30434231.0   5119954.0   6303371.0
2003-09-07 20:20:00   6931547.0  23795055.0   3817257.0   7016878.0
2003-09-07 20:30:00   7191846.0  27884459.0   3467106.0   6755673.0
2003-09-07 20:40:00   6496644.0  29650677.0   3444472.0   6628083.0
2003-09-07 20:50:00   5357749.0  30596991.0   3674740.0   6397438.0
2003-09-07 21:00:00   4954037.0  32431427.0   3298414.0   5899602.0
2003-09-07 21:10:00   6607095.0  31753221.0   5221478.0   4921226.0
2003-09-07 21:20:00   7761014.0  32440126.0   5270735.0   5113812.0
2003-09-07 21:30:00   8278082.0  32339499.0   4486743.0   3844078.0
2003-09-07 21:40:00  14819066.0  33316562.0   4420518.0   3865739.0
2003-09-07 21:50:00  15187761.0  32930008.0   5654167.0   6380081.0
2003-09-07 22:00:00   8610814.0  29580144.0   7372781.0   5177476.0
2003-09-07 22:10:00  14989717.0  33503534.0   5279433.0   4335068.0
2003-09-07 22:20:00   8255256.0  37512167.0   9005777.0   5046527.0
2003-09-07 22:30:00  11540450.0  42011173.0   9201234.0   4804330.0
2003-09-07 22:40:00   9113018.0  34696766.0   6105271.0   4773575.0
2003-09-07 22:50:00  10189755.0  32945338.0   6104020.0   4200284.0
2003-09-07 23:00:00  10481787.0  33946248.0   6716219.0   3387668.0
2003-09-07 23:10:00   8849096.0  33461807.0   5866138.0   3786793.0
2003-09-07 23:20:00   9776675.0  31474607.0   5874654.0  11277465.0
2003-09-07 23:30:00   9144621.0  32117262.0   5762691.0   7154577.0
2003-09-07 23:40:00   8802106.0  29932510.0   5279285.0   5950898.0
```

```
2003-09-07 23:50:00   8716795.6  22660870.0   6240626.4   5657380.6

                      ATLA-IPLS   ATLA-KSCY   ATLA-LOSA   ATLA-NYCM  \
2003-09-01 00:00:00  21756443.0  10792818.0  14220940.0  25014340.0
2003-09-01 00:10:00  24497174.0   8623734.0  15695839.0  36788680.0
2003-09-01 00:20:00  23254392.0   7882377.0  16176022.0  31682355.0
2003-09-01 00:30:00  28747761.0   9102603.0  16200072.0  27472465.0
2003-09-01 00:40:00  35642229.0   9104036.0  12279530.0  29171205.0
2003-09-01 00:50:00  23691423.0  12097067.0  15160907.0  35705296.0
2003-09-01 01:00:00  29599650.0  14222361.0  16047109.0  31734558.0
2003-09-01 01:10:00  22085051.0  12169131.0  12351694.0  37517103.0
2003-09-01 01:20:00  30503423.0   9387983.0  14167375.0  36373669.0
2003-09-01 01:30:00  25435393.0   9935391.0  14640947.0  34719841.0
2003-09-01 01:40:00  29436475.0   7966885.0  13286314.0  32630768.0
2003-09-01 01:50:00  29466210.0   7112299.0  14683964.0  33551931.0
2003-09-01 02:00:00  24506432.0   6387421.0  14713899.0  28826257.0
2003-09-01 02:10:00  27860028.0   7116898.0  15008541.0  28633753.0
2003-09-01 02:20:00  27475432.0   9609178.0  15239638.0  25608032.0
2003-09-01 02:30:00  29606534.0   6600804.0  16563554.0  27022739.0
2003-09-01 02:40:00  31959958.0   8368112.0  17043800.0  26830609.0
2003-09-01 02:50:00  26756472.0   9561098.0  16533364.0  24065162.0
2003-09-01 03:00:00  28867579.0   5956872.0  13811222.0  21207074.0
2003-09-01 03:10:00  27306642.0   7078270.0  17369912.0  25567713.0
2003-09-01 03:20:00  25344287.0   6689851.0  18321688.0  23303418.0
2003-09-01 03:30:00  27277718.0   8110981.0  15069516.0  27067467.0
2003-09-01 03:40:00  28094690.0   7661749.0  15997935.0  27524269.0
2003-09-01 03:50:00  29014447.0   8841130.0  17932116.0  23781421.0
2003-09-01 04:00:00  27479647.0   8310171.0  17462697.0  23585860.0
2003-09-01 04:10:00  31115283.0   6910929.0  16287357.0  28149338.0
2003-09-01 04:20:00  26132982.0   7601269.0  16118521.0  24127447.0
2003-09-01 04:30:00  28766051.0   6214227.0  13409383.0  24129985.0
2003-09-01 04:40:00  33204543.0   8327256.0  14439401.0  22186961.0
2003-09-01 04:50:00  32455550.0   7191669.0  18008152.0  25333790.0
...                         ...         ...         ...         ...
2003-09-07 19:00:00  21779165.0   6311288.0  11155492.0  18033950.0
2003-09-07 19:10:00  18499485.0   8006986.0  11997277.0  25084019.0
2003-09-07 19:20:00  14239089.0   4111967.0  10789154.0  22640113.0
2003-09-07 19:30:00  19491095.0   5876478.0   6629376.0  21369079.0
2003-09-07 19:40:00  18611592.0   7434331.0   7400496.0  20223097.0
2003-09-07 19:50:00  15359764.0   5219028.0  12511375.0  22306629.0
2003-09-07 20:00:00  17389950.0   6383892.0  11687799.0  19653060.0
2003-09-07 20:10:00  19953012.0   7354133.0  14431223.0  23834905.0
2003-09-07 20:20:00  14907457.0   7846717.0  15196580.0  28301505.0
2003-09-07 20:30:00  16396674.0   6702042.0  19101974.0  27080261.0
2003-09-07 20:40:00  22430750.0   6189345.0  12527947.0  22819828.0
2003-09-07 20:50:00  22828898.0   7007960.0   8107466.0  26228283.0
2003-09-07 21:00:00  26547060.0   6186291.0   7817750.0  28516072.0
2003-09-07 21:10:00  25228887.0  13127619.0   8187446.0  30109970.0
```

```
2003-09-07 21:20:00  26707406.0   8001751.0   7453185.0  24356401.0
2003-09-07 21:30:00  22387140.0   7092630.0  16863592.0  24445525.0
2003-09-07 21:40:00  25236624.0   7605333.0   9089740.0  26039854.0
2003-09-07 21:50:00  23179401.0   8717879.0  10075222.0  26071536.0
2003-09-07 22:00:00  25893394.0   7733513.0  15213286.0  22927205.0
2003-09-07 22:10:00  22895219.0   7275602.0   9852731.0  26838423.0
2003-09-07 22:20:00  15704701.0   7804769.0  14232313.0  24911722.0
2003-09-07 22:30:00  15001822.0   8086615.0  21878211.0  24476655.0
2003-09-07 22:40:00  18875088.0   9290472.0  15394337.0  24958629.0
2003-09-07 22:50:00  15010103.0  10511764.0  15969744.0  21748865.0
2003-09-07 23:00:00  19046909.0  11253778.0  25786657.0  22193749.0
2003-09-07 23:10:00  19097140.0  10561532.0  26092040.0  28640962.0
2003-09-07 23:20:00  14314837.0   9106198.0  26412752.0  26168288.0
2003-09-07 23:30:00  17771350.0  10149256.0  29501669.0  25998158.0
2003-09-07 23:40:00  20222187.0  10636832.0  19613671.0  26124024.0
2003-09-07 23:50:00  17406086.0   8808588.5  15962917.0  18367639.0

                      ATLA-SNVA   ATLA-STTL    ...         WASH-CHIN  \
2003-09-01 00:00:00  13677284.0  10591345.0    ...        53296727.0
2003-09-01 00:10:00   5607086.0  10714795.0    ...        68413060.0
2003-09-01 00:20:00   6354657.0  12205515.0    ...        67969461.0
2003-09-01 00:30:00   9402609.0  10934084.0    ...        66616097.0
2003-09-01 00:40:00   7624924.0  11327807.0    ...        66797282.0
2003-09-01 00:50:00   7139036.0  10426541.0    ...        63664403.0
2003-09-01 01:00:00   6619160.0  10412081.0    ...        58286708.0
2003-09-01 01:10:00   6039646.0  10705797.0    ...        62172268.0
2003-09-01 01:20:00   4669164.0  10228338.0    ...        68251716.0
2003-09-01 01:30:00   9286745.0  10992822.0    ...        64817476.0
2003-09-01 01:40:00   4672156.0  11502073.0    ...        64994636.0
2003-09-01 01:50:00   5489845.0  10176607.0    ...        59458031.0
2003-09-01 02:00:00  13888828.0  10287707.0    ...        67896792.0
2003-09-01 02:10:00   6940811.0  10069618.0    ...        66870509.0
2003-09-01 02:20:00   7114755.0  10272844.0    ...        65507693.0
2003-09-01 02:30:00   9453120.0   9772422.0    ...        74129601.0
2003-09-01 02:40:00   6416275.0  10026158.0    ...        81327390.0
2003-09-01 02:50:00   4357088.0  10586196.0    ...        78942615.0
2003-09-01 03:00:00   4944809.0  10124193.0    ...        77172158.0
2003-09-01 03:10:00   5297942.0  10639443.0    ...        77670762.0
2003-09-01 03:20:00   6861689.0  12481133.0    ...        69047624.0
2003-09-01 03:30:00   8786181.0  10859930.0    ...        71841607.0
2003-09-01 03:40:00   5609763.0  10270395.0    ...        74965284.0
2003-09-01 03:50:00   7360153.0  10921262.0    ...        77758286.0
2003-09-01 04:00:00  15090883.0  10835773.0    ...        73908303.0
2003-09-01 04:10:00  10468197.0  11089254.0    ...        71201166.0
2003-09-01 04:20:00  11523419.0  12239221.0    ...        70570974.0
2003-09-01 04:30:00  11591146.0  10834976.0    ...        71928476.0
2003-09-01 04:40:00   9528480.0  10523968.0    ...        61813637.0
2003-09-01 04:50:00   9484484.0  11435195.0    ...        56336087.0
```

```
...                         ...         ...       ...           ...
2003-09-07 19:00:00   4585239.0   8516498.0       ...    56008988.0
2003-09-07 19:10:00   5719757.0   7564477.0       ...    57253777.0
2003-09-07 19:20:00   5285885.0   7490111.0       ...    53269268.0
2003-09-07 19:30:00   8788618.0   7761440.0       ...    57667375.0
2003-09-07 19:40:00   6338335.0   8198256.0       ...    52629442.0
2003-09-07 19:50:00   5788810.0   7892931.0       ...    50079454.0
2003-09-07 20:00:00   5505913.0   6958396.0       ...    58008231.0
2003-09-07 20:10:00   6415492.0   7933142.0       ...    61595649.0
2003-09-07 20:20:00   5284332.0   7856418.0       ...    56119538.0
2003-09-07 20:30:00   9448919.0   8100297.0       ...    58841177.0
2003-09-07 20:40:00   8907769.0   7913160.0       ...    62444994.0
2003-09-07 20:50:00   7430860.0   7806629.0       ...    66059435.0
2003-09-07 21:00:00   8451747.0   8257688.0       ...    63440574.0
2003-09-07 21:10:00   9195010.0   7962118.0       ...    60127070.0
2003-09-07 21:20:00  11359020.0   7632903.0       ...    58020393.0
2003-09-07 21:30:00  12835979.0   8244838.0       ...    63038589.0
2003-09-07 21:40:00  11627709.0   8476405.0       ...    68951904.0
2003-09-07 21:50:00  10663149.0  10448948.0       ...    67489771.0
2003-09-07 22:00:00   9793167.0   8707757.0       ...    70290052.0
2003-09-07 22:10:00  10038563.0   7855078.0       ...    73249215.0
2003-09-07 22:20:00   8851054.0   7861153.0       ...    68214855.0
2003-09-07 22:30:00  10168798.0   8881080.0       ...    71505364.0
2003-09-07 22:40:00   7594436.0   9482891.0       ...    67735084.0
2003-09-07 22:50:00  10175619.0   9119160.0       ...    71974381.0
2003-09-07 23:00:00   8188090.0   9566768.0       ...    64323386.0
2003-09-07 23:10:00   8343867.0   8820650.0       ...    65925313.0
2003-09-07 23:20:00   8638782.0   9193717.0       ...    70075490.0
2003-09-07 23:30:00  11343171.0   9423042.0       ...    68544458.0
2003-09-07 23:40:00   8732768.0   8217873.0       ...    65087776.0
2003-09-07 23:50:00   7767967.3   7470650.1       ...    65599891.0

                      WASH-DNVR   WASH-HSTN   WASH-IPLS   WASH-KSCY  \
2003-09-01 00:00:00  18724766.0  12238893.0  52782009.0  12836459.0
2003-09-01 00:10:00  28522606.0  11377094.0  60006620.0  12556471.0
2003-09-01 00:20:00  37073856.0  15680615.0  61484233.0  16318506.0
2003-09-01 00:30:00  43019246.0  12726958.0  64027333.0  16394673.0
2003-09-01 00:40:00  40408580.0  11733121.0  54541962.0  16769259.0
2003-09-01 00:50:00  37653260.0  13189909.0  58056897.0  18505687.0
2003-09-01 01:00:00  30293795.0  15201058.0  62827807.0  17116749.0
2003-09-01 01:10:00  23985085.0  14492978.0  65236857.0  17821662.0
2003-09-01 01:20:00  36661473.0  15587404.0  57840291.0  15055479.0
2003-09-01 01:30:00  40307899.0  11192637.0  56222991.0  16064173.0
2003-09-01 01:40:00  31251456.0  12583005.0  60510052.0  16678774.0
2003-09-01 01:50:00  32890232.0  13929604.0  58135473.0  17214940.0
2003-09-01 02:00:00  34965289.0  17218126.0  61923336.0  20034242.0
2003-09-01 02:10:00  33257337.0  18659363.0  67085271.0  18990421.0
2003-09-01 02:20:00  35473142.0  18280908.0  60352455.0  20163554.0
```

```
2003-09-01 02:30:00  34720792.0  19832860.0  58473573.0  20364539.0
2003-09-01 02:40:00  34509148.0  13496515.0  65651235.0  18303701.0
2003-09-01 02:50:00  37810387.0  13736059.0  63964710.0  19859609.0
2003-09-01 03:00:00  23401984.0  13645410.0  68376174.0  17379740.0
2003-09-01 03:10:00  25194390.0  14663308.0  71332611.0  16605148.0
2003-09-01 03:20:00  31416665.0  19962279.0  67695567.0  17097792.0
2003-09-01 03:30:00  38792234.0  15227079.0  64279393.0  17904445.0
2003-09-01 03:40:00  37912498.0  13378723.0  68911615.0  17786671.0
2003-09-01 03:50:00  35559608.0  12721115.0  74285111.0  15936781.0
2003-09-01 04:00:00  35276563.0  13138714.0  77620251.0  17972364.0
2003-09-01 04:10:00  34916285.0  18589199.0  79366451.0  19677460.0
2003-09-01 04:20:00  33564858.0  14718878.0  80856867.0  19586501.0
2003-09-01 04:30:00  30023351.0  16371112.0  79269129.0  19980312.0
2003-09-01 04:40:00  29153384.0  10094092.0  81576670.0  18373943.0
2003-09-01 04:50:00  30302265.0  16094102.0  78596140.0  17612306.0
...                         ...         ...         ...         ...
2003-09-07 19:00:00  22724031.0   9287522.0  45535447.0  12739190.0
2003-09-07 19:10:00  20882759.0   8821192.0  45531667.0  10933572.0
2003-09-07 19:20:00  22048509.0  13240807.0  49048450.0  12704103.0
2003-09-07 19:30:00  24258640.0  12572454.0  48091482.0  12886328.0
2003-09-07 19:40:00  29789736.0  13381971.0  47551147.0  18104673.0
2003-09-07 19:50:00  28081309.0  12567254.0  42338696.0  18888130.0
2003-09-07 20:00:00  29346723.0  12078458.0  45319625.0  15287306.0
2003-09-07 20:10:00  24489486.0  14753447.0  45614683.0  14829584.0
2003-09-07 20:20:00  28258142.0  17298211.0  51085330.0  16212472.0
2003-09-07 20:30:00  25497974.0  13684016.0  50974989.0  15688904.0
2003-09-07 20:40:00  30059845.0  13154633.0  53908699.0  13143840.0
2003-09-07 20:50:00  32247958.0  15349346.0  54245267.0  12548658.0
2003-09-07 21:00:00  21661112.0  13995628.0  54552939.0  12812705.0
2003-09-07 21:10:00  21746823.0  15838981.0  42949121.0  12776961.0
2003-09-07 21:20:00  27109821.0  16812313.0  43626340.0  12856479.0
2003-09-07 21:30:00  29168051.0  12166286.0  42344355.0  16927493.0
2003-09-07 21:40:00  33509017.0   9506685.0  49607079.0  15978986.0
2003-09-07 21:50:00  30831069.0  10956856.0  53294442.0  16890886.0
2003-09-07 22:00:00  33182446.0  12203390.0  39483143.0  15446733.0
2003-09-07 22:10:00  32363598.0  11795037.0  48370006.0  13594915.0
2003-09-07 22:20:00  30935501.0  15039492.0  51477661.0  12114461.0
2003-09-07 22:30:00  28905415.0  12563476.0  52685141.0  14658137.0
2003-09-07 22:40:00  28800830.0  17291394.0  49308368.0  17484422.0
2003-09-07 22:50:00  33295923.0  12692831.0  50739873.0  21504695.0
2003-09-07 23:00:00  24645263.0  13504904.0  52455210.0  19114819.0
2003-09-07 23:10:00  21751316.0  11058944.0  58591021.0  17137907.0
2003-09-07 23:20:00  29126443.0  12667321.0  54571764.0  15383038.0
2003-09-07 23:30:00  27817836.0  15892668.0  50326213.0  12098328.0
2003-09-07 23:40:00  28836922.0  11075541.0  52574692.0  11933512.0
2003-09-07 23:50:00  25862152.0  11673804.0  60086953.0  11851656.0

                      WASH-LOSA   WASH-NYCM   WASH-SNVA   WASH-STTL  \
2003-09-01 00:00:00  31460190.0  105796930.0  13756184.0  13582945.0
2003-09-01 00:10:00  32450393.0   70665497.0  13968786.0  16144471.0
2003-09-01 00:20:00  33768245.0   71577084.0  13938533.0  14959708.0
2003-09-01 00:30:00  33440318.0   79682647.0  16212806.0  16425845.0
2003-09-01 00:40:00  33927515.0   81480788.0  16757707.0  15158825.0
2003-09-01 00:50:00  32377995.0  105472620.0  16170743.0  14282972.0
2003-09-01 01:00:00  29311889.0   78615122.0  17669077.0  16421471.0
2003-09-01 01:10:00  32904086.0   86022060.0  16735178.0  20557085.0
2003-09-01 01:20:00  33461380.0   82169945.0  15783877.0  19195575.0
2003-09-01 01:30:00  32535261.0   96716354.0  14813943.0  19581037.0
2003-09-01 01:40:00  28535178.0  110099130.0  12866245.0  20499171.0
2003-09-01 01:50:00  28319036.0   81321006.0  12514823.0  22032745.0
2003-09-01 02:00:00  28992843.0   87758864.0  13083396.0  20490276.0
2003-09-01 02:10:00  26738059.0   85719542.0  16380609.0  23213334.0
2003-09-01 02:20:00  36791149.0   89979179.0  19247125.0  22629123.0
2003-09-01 02:30:00  38214954.0   74917010.0  18759730.0  20846535.0
2003-09-01 02:40:00  29538309.0   80767619.0  20755012.0  18598317.0
2003-09-01 02:50:00  32367804.0   78420778.0  22465075.0  17091704.0
2003-09-01 03:00:00  36225291.0  109322620.0  24684493.0  16508101.0
2003-09-01 03:10:00  37738866.0   99426914.0  26672590.0  19232286.0
2003-09-01 03:20:00  35576530.0   94434993.0  26469234.0  19007746.0
2003-09-01 03:30:00  29252599.0   89358278.0  22773366.0  20368470.0
2003-09-01 03:40:00  34265582.0  114804930.0  16912557.0  20130887.0
2003-09-01 03:50:00  33562949.0   99833526.0  17476002.0  20872647.0
2003-09-01 04:00:00  32455609.0  102758620.0  19537415.0  21316664.0
2003-09-01 04:10:00  33531609.0  119641760.0  22281459.0  21804663.0
2003-09-01 04:20:00  35992956.0  122647880.0  21429104.0  20848513.0
2003-09-01 04:30:00  38607893.0  107333240.0  17406190.0  20151076.0
2003-09-01 04:40:00  33326104.0  112031820.0  17643968.0  20540767.0
2003-09-01 04:50:00  36334953.0  129997660.0  17325631.0  18356338.0
...                         ...          ...         ...         ...
2003-09-07 19:00:00  14294344.0   69282596.0  11543843.0  18799826.0
2003-09-07 19:10:00  23454470.0   91708496.0  15622473.0  19979163.0
2003-09-07 19:20:00  23109376.0   76975716.0  13641494.0  20549597.0
2003-09-07 19:30:00  22815464.0   80474247.0  11646187.0  20466986.0
2003-09-07 19:40:00  22587720.0   74431040.0  10461321.0  15869656.0
2003-09-07 19:50:00  22277638.0   70714341.0   9636593.0  14295295.0
2003-09-07 20:00:00  25243028.0  115578110.0  13812280.0  15965280.0
2003-09-07 20:10:00  26929308.0   92763309.0  14631122.0  18653648.0
2003-09-07 20:20:00  25662498.0   76539842.0  12779750.0  16869328.0
2003-09-07 20:30:00  25059034.0   73883704.0  16838171.0  16764719.0
2003-09-07 20:40:00  28568819.0   73391126.0  19209880.0  15222354.0
2003-09-07 20:50:00  22756530.0   71637735.0  19571655.0  14888930.0
2003-09-07 21:00:00  22475372.0   68865349.0  20444568.0  16644783.0
2003-09-07 21:10:00  21691398.0   72404743.0  18513584.0  21198006.0
2003-09-07 21:20:00  21086136.0   62868895.0  19814599.0  18611813.0
2003-09-07 21:30:00  21436715.0   66175210.0  21788777.0  18985025.0
2003-09-07 21:40:00  19531118.0   70971966.0  20666630.0  16979654.0
```

```
2003-09-07 21:50:00  21234144.0   70894572.0  20532931.0  15980801.0
2003-09-07 22:00:00  21754051.0   74068192.0  20521312.0  14314325.0
2003-09-07 22:10:00  21341302.0   77246108.0  21322688.0  18313070.0
2003-09-07 22:20:00  21687938.0   62929140.0  18277323.0  15960957.0
2003-09-07 22:30:00  26157166.0   65771209.0  20345687.0  17599724.0
2003-09-07 22:40:00  25126968.0   68864540.0  20840934.0  16037346.0
2003-09-07 22:50:00  25448516.0   73626140.0  17109318.0  16448103.0
2003-09-07 23:00:00  26312827.0   86447406.0  16455697.0  17795684.0
2003-09-07 23:10:00  24297674.0   83293655.0  17329425.0  20865535.0
2003-09-07 23:20:00  25238842.0   70015955.0  16526455.0  16881206.0
2003-09-07 23:30:00  27689197.0   73553203.0  18022288.0  18471915.0
2003-09-07 23:40:00  31632344.0   81693475.0  16677568.0  16766967.0
2003-09-07 23:50:00  30979811.0   73577193.0  19167646.0  19402758.0

                       WASH-WASH
2003-09-01 00:00:00  120384980.0
2003-09-01 00:10:00  135679630.0
2003-09-01 00:20:00  126175780.0
2003-09-01 00:30:00  112891500.0
2003-09-01 00:40:00  123140310.0
2003-09-01 00:50:00  113083050.0
2003-09-01 01:00:00  125216470.0
2003-09-01 01:10:00  120186790.0
2003-09-01 01:20:00  116369030.0
2003-09-01 01:30:00  121708940.0
2003-09-01 01:40:00  125549510.0
2003-09-01 01:50:00  126531040.0
2003-09-01 02:00:00  130922860.0
2003-09-01 02:10:00  132442070.0
2003-09-01 02:20:00  139187180.0
2003-09-01 02:30:00  133343820.0
2003-09-01 02:40:00  142018400.0
2003-09-01 02:50:00  147258260.0
2003-09-01 03:00:00  150228660.0
2003-09-01 03:10:00  143673670.0
2003-09-01 03:20:00  146708640.0
2003-09-01 03:30:00  142494930.0
2003-09-01 03:40:00  132745340.0
2003-09-01 03:50:00  135520370.0
2003-09-01 04:00:00  136599240.0
2003-09-01 04:10:00  136553280.0
2003-09-01 04:20:00  134846690.0
2003-09-01 04:30:00  129035150.0
2003-09-01 04:40:00  130436100.0
2003-09-01 04:50:00  139999310.0
...                          ...
2003-09-07 19:00:00   92615400.0
2003-09-07 19:10:00   97569918.0
```

```
         2003-09-07 19:20:00   90794878.0
         2003-09-07 19:30:00   96209878.0
         2003-09-07 19:40:00   90844693.0
         2003-09-07 19:50:00   98665658.0
         2003-09-07 20:00:00   93786912.0
         2003-09-07 20:10:00  103585960.0
         2003-09-07 20:20:00  111043730.0
         2003-09-07 20:30:00  105928030.0
         2003-09-07 20:40:00  102967340.0
         2003-09-07 20:50:00  106496810.0
         2003-09-07 21:00:00  114167630.0
         2003-09-07 21:10:00  116108220.0
         2003-09-07 21:20:00  117503140.0
         2003-09-07 21:30:00  118484880.0
         2003-09-07 21:40:00  133833880.0
         2003-09-07 21:50:00  130826570.0
         2003-09-07 22:00:00  137975450.0
         2003-09-07 22:10:00  122758600.0
         2003-09-07 22:20:00  113300780.0
         2003-09-07 22:30:00  115715010.0
         2003-09-07 22:40:00  111927980.0
         2003-09-07 22:50:00  111120880.0
         2003-09-07 23:00:00  101430510.0
         2003-09-07 23:10:00  123125390.0
         2003-09-07 23:20:00  142106800.0
         2003-09-07 23:30:00  127918530.0
         2003-09-07 23:40:00  138180630.0
         2003-09-07 23:50:00  137288810.0

         [1008 rows x 121 columns]
```

```
In [66]: Atraf.shape
```

```
Out[66]: (1008, 121)
```

As we would expect, our traffic matrix has rank 121:

```
In [67]: np.linalg.matrix_rank(Atraf)
```

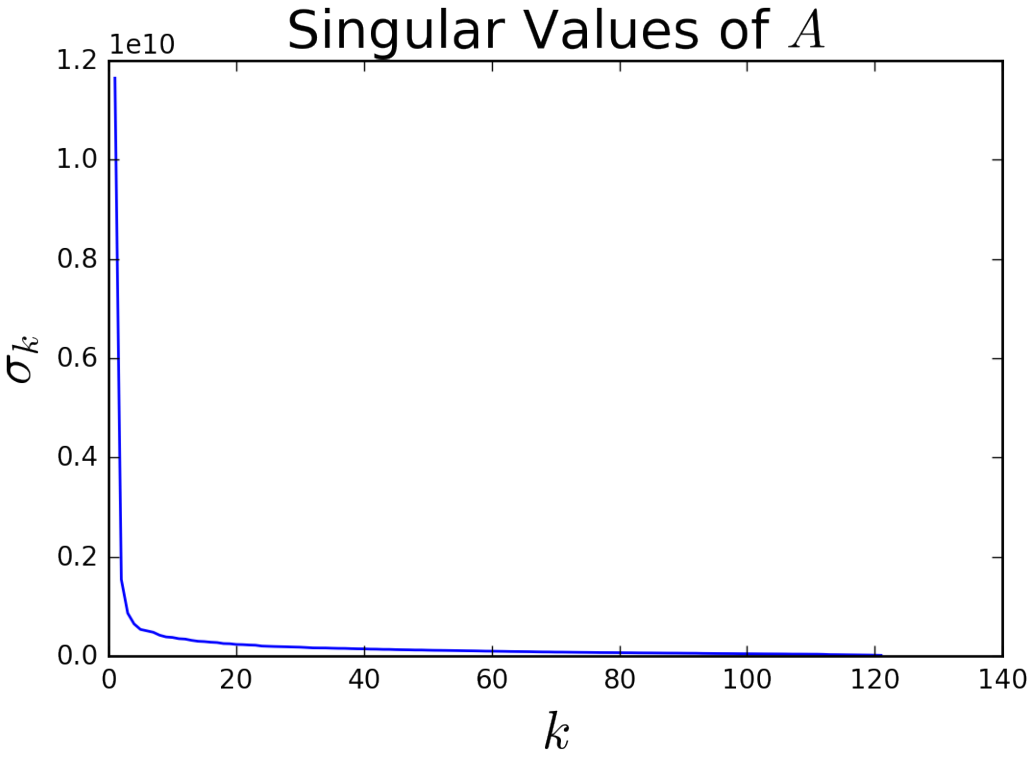
```
Out[67]: 121
```

However – perhaps it has low **effective** rank.
The `numpy` routine for computing SVD is `np.linalg.svd`:

```
In [68]: u,s,vt = np.linalg.svd(Atraf)
```

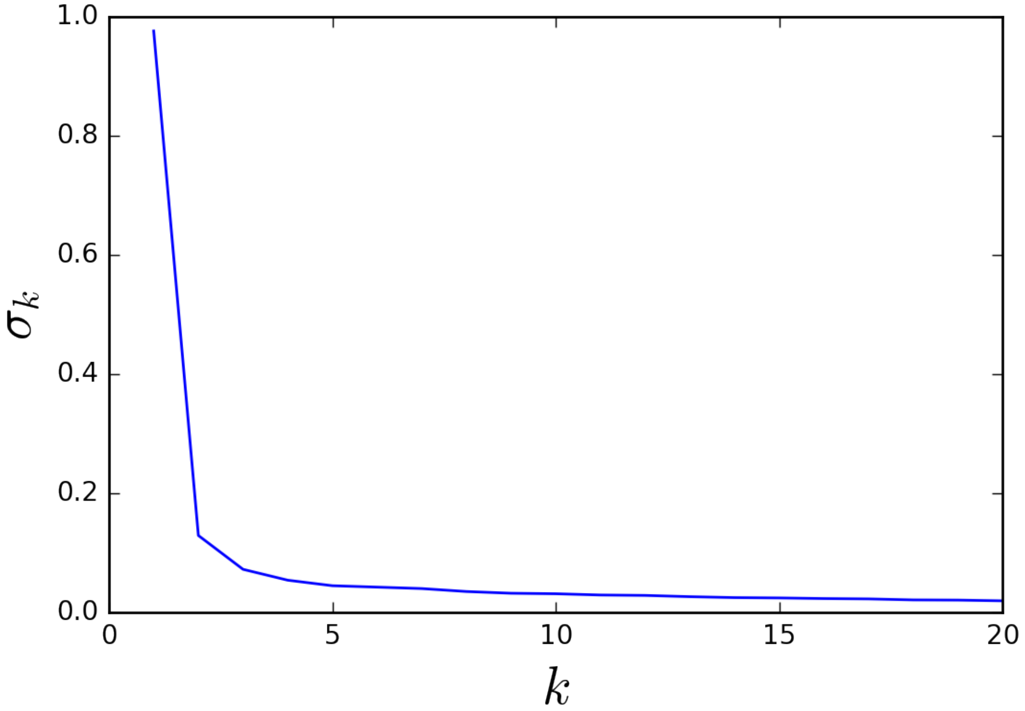
Now let's look at the singular values of `Atraf` to see if it can be usefully approximated as a low-rank matrix:

```
In [69]: 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 $A$',size=20)
```



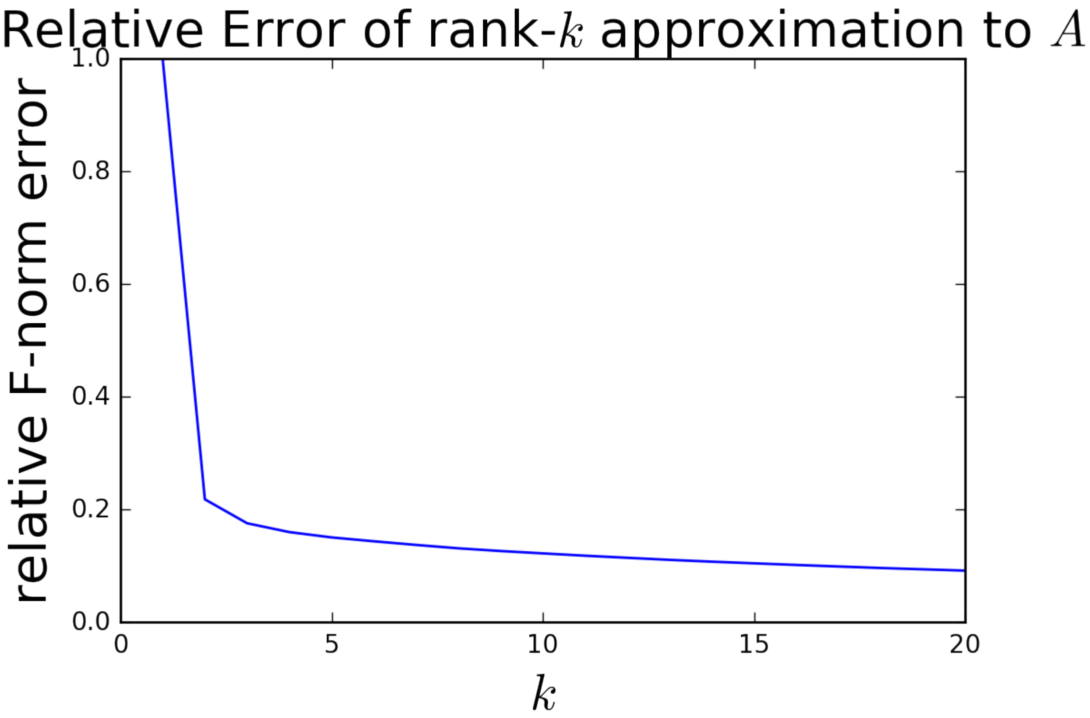
Zooming in:

```
In [70]: fig = plt.figure(figsize=(6,4))
         Anorm = np.linalg.norm(Atraf)
         plt.plot(range(1,21),s[0:20]/Anorm)
         plt.xlim([0,20])
         plt.xlabel(r'$k$',size=20)
         _ = plt.ylabel(r'$\sigma_k$',size=20)
```



Let's use the formula to compute the relative error of a rank-k approximation to A:

```
In [71]: fig = plt.figure(figsize=(6,4))
         Anorm = np.linalg.norm(Atraf)
         err = np.cumsum(s[::-1]**2)
         err = np.sqrt(err[::-1])
         plt.plot(range(1,21),err[:20]/Anorm)
         plt.xlim([0,20])
         plt.xlabel(r'$k$',size=20)
         plt.ylabel(r'relative F-norm error',size=20)
         _ = plt.title(r'Relative Error of rank-$k$ approximation to $A$',size=20)
```



```
In [72]: err[0:20]/Anorm

Out[72]: array([ 1.        ,  0.21778461,  0.17530185,  0.15964863,  0.15018706,
                 0.14332383,  0.13687439,  0.1308678 ,  0.12606339,  0.12187111,
                 0.11775775,  0.11405379,  0.11040251,  0.10717555,  0.10423266,
                 0.10133916,  0.09861588,  0.09595619,  0.09361865,  0.09129217])
```

So we are down to 9% relative error using a rank-20 approximation to A.

So instead of storing $mn =$ (1008 · 121) = 121,968 values, we only need to store $k(m+n)$ = 20 · (1008 + 121) = 22,580 values, which is a 81% reduction in size.

1.7 Low Effective Rank is Common

In practice many datasets have low effective rank. Here are some more examples.

Likes on Facebook.

Here, the matrices are

1. Number of likes: Timebins × Users
2. Number of likes: Users × Page Categories
3. Entropy of likes across categories: Timebins × Users

Source: [Viswinath et al.]

Social Media Activity.

Here, the matrices are

1. Number of Yelp reviews: Timebins × Users
2. Number of Yelp reviews: Users × Yelp Categories
3. Number of Tweets: Users × Topic Categories

Source: [Viswinath et al.]

User preferences over items.

Example: the Netflix prize worked with partially-observed matrices like this:

$$\begin{bmatrix} & & & \vdots & & & \\ & & 3 & 2 & & 1 & \\ & 1 & & 1 & & & \\ \cdots & & 2 & & 4 & & \cdots \\ & 5 & 5 & & 4 & & \\ & 1 & & & 1 & 5 & \\ & & & \vdots & & & \end{bmatrix}$$

Where the rows correspond to users, the columns to movies, and the entries are ratings.

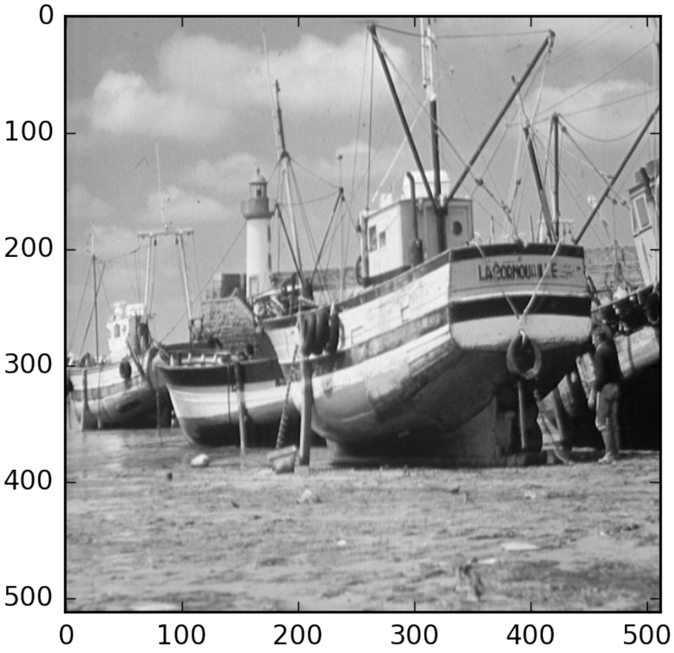
Although the problem matrix was of size 500,000 × 18,000, the winning approach modeled the matrix as having **rank 20 to 40.**

Images.

Image data often shows low effective rank.

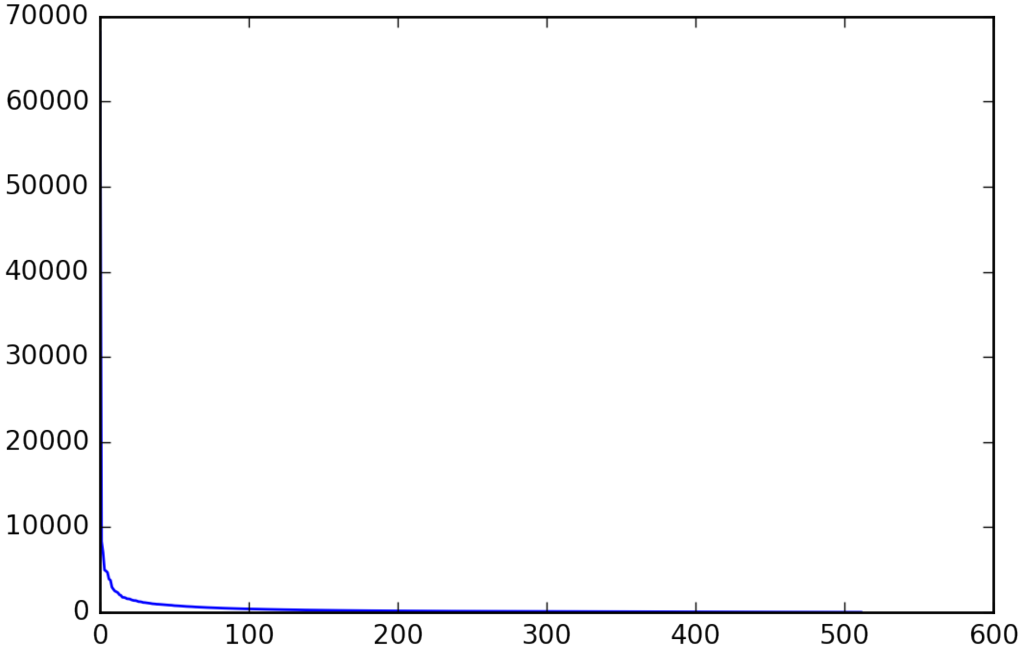
For example, here is an original photo:

```
In [73]: boat = np.loadtxt('data/images/boat/boat.dat')
         import matplotlib.cm as cm
         plt.figure()
         _ = plt.imshow(boat,cmap = cm.Greys_r)
```



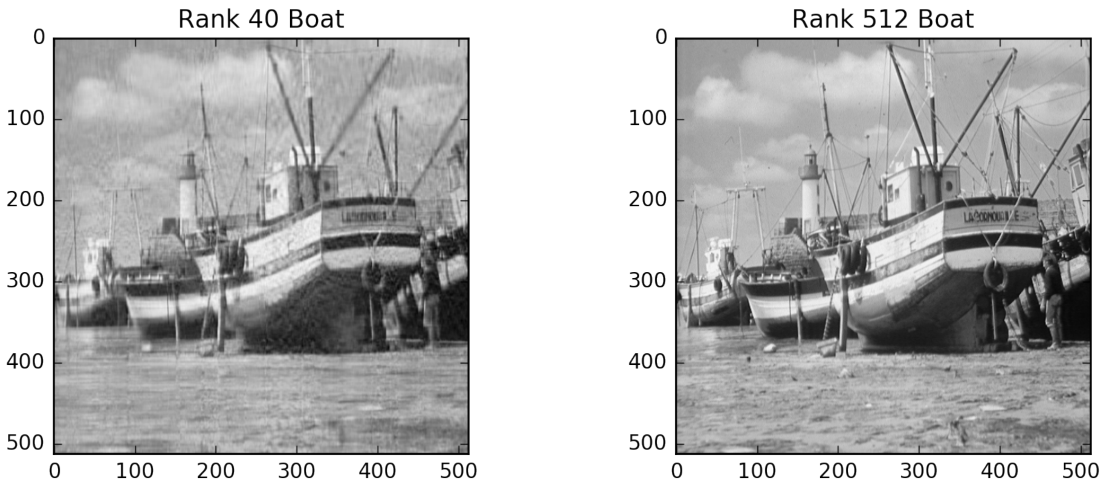
Let’s look at its spectrum:

```
In [74]: u,s,vt=np.linalg.svd(boat,full_matrices=False)
         _ = plt.plot(s)
```



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 [75]: # construct a rank-40 version of the boat
         scopy = s.copy()
         scopy[40:]=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 40 Boat')
         plt.subplot(1,2,2)
         plt.imshow(boat,cmap = cm.Greys_r)
         plt.title('Rank 512 Boat')
         _ = plt.subplots_adjust(wspace=0.5)
```



1.8 Interpretations of Low Effective Rank

How can we understand the low-effective-rank phenomenon in general?

There are two helpful interpretations:

1. Common Patterns
2. Latent Factors

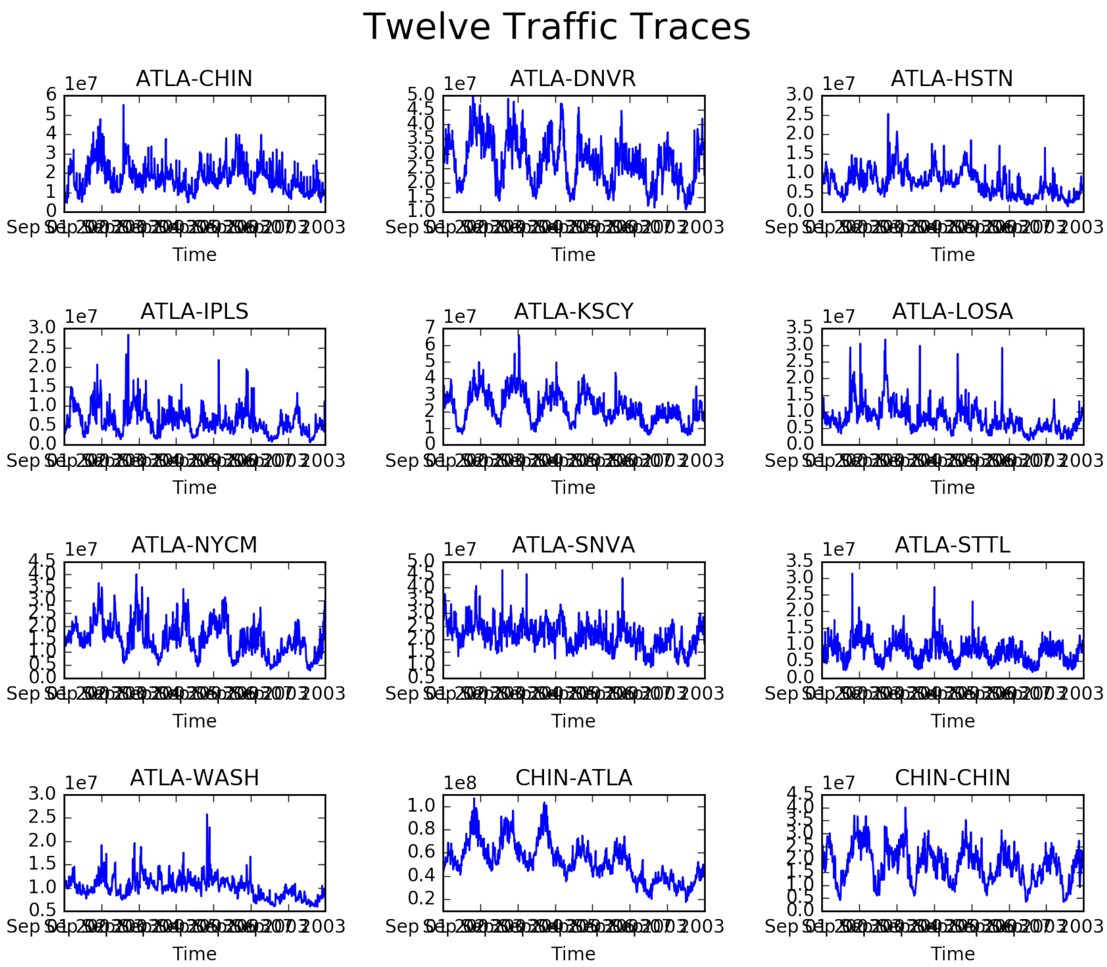
1.8.1 Common Patterns.

The first interpretation of low-rank behavior is in answering the question:

"What is the strongest pattern in the data?"

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

In this interpretation, we think of each column of A as a combination of the columns of U'.



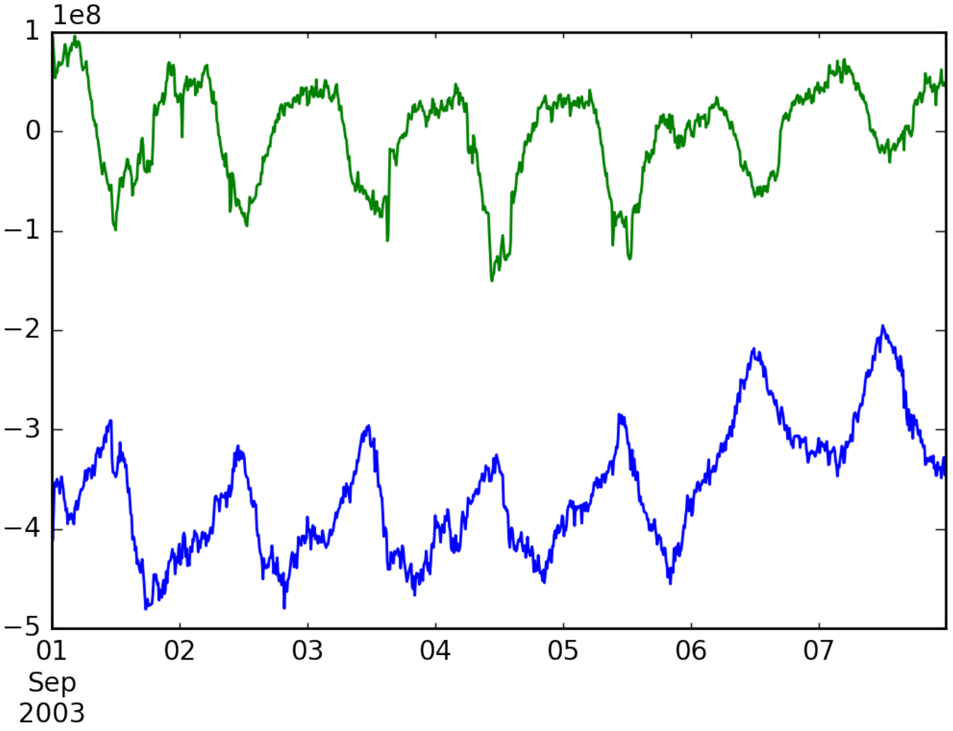
Let's use as our example $\mathbf{a}_1$, the first column of A.
This happens to be the ATLA-CHIN flow.
The equation above tells us that

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

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.

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

```
In [77]: 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',fre
         uframe[0].plot()
         _ = uframe[1].plot()
```



1.8.2 Latent Factors.

The next interpretation of low-rank behavior is that it exposes "latent factors" that describe the data.

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

In this interpretation, we think of each element of A as the inner product of a row of $U'\Sigma'$ and a row of V'.

Let's say we are working with a matrix of users and items.

In particular, let item be movies and matrix entries be ratings, as in the Netflix prize.

Recall the structure from a previous slide:

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

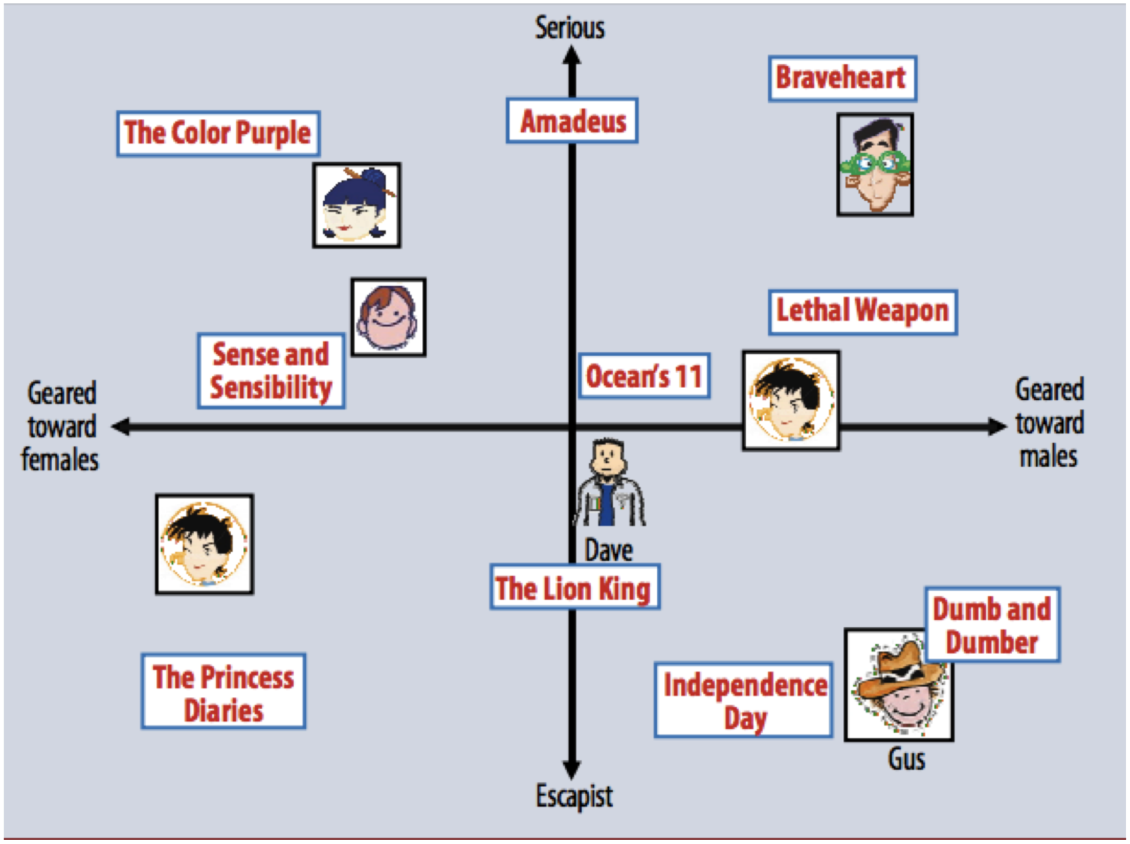
Then the rating that a user gives a movie is the inner product of a k element vector that corresponds to the user, and a k element vector that corresponds to the movie.

We can therefore think of each user's preferences as being captured by point in $\mathbb{R}^k$. This is a **latent factor.**

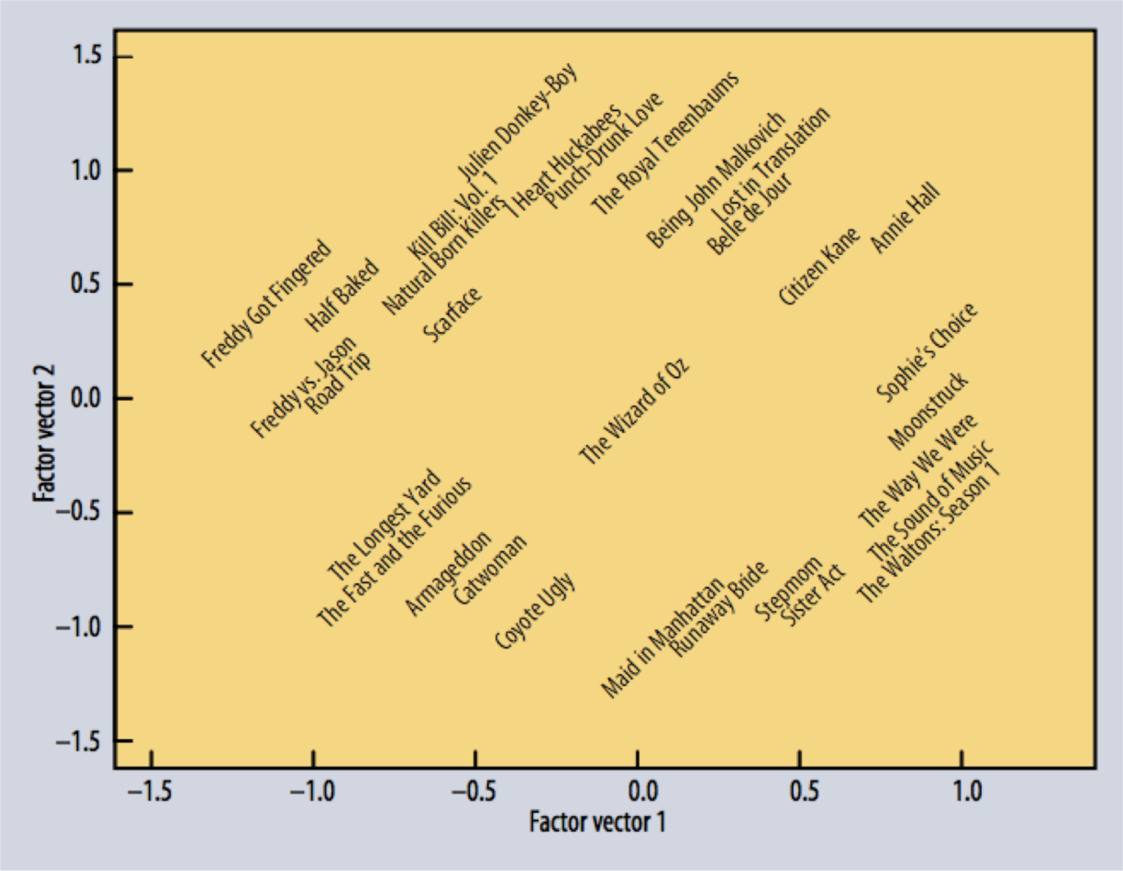
The value in this comes from the summarization of user preferences, and the predictive power it gives.

For example, the winning entry in the Netflix prize competition modeled user preferences with a 20-element latent factor.

The remarkable thing is that a person's preferences for all 18,000 movies can be captured in a 20-element vector!



Source: Koren et al, IEEE Computer, 2009



Source: Koren et al, IEEE Computer, 2009