Lecture #14: Low Rank Approximation

Today, let's start with a thought experiment

Q: Suppose I start with a rank one matrix

$$\underset{n\times n}{A} = \underset{\text{unit vectors}}{u v^T}$$

and add random numbers in $[-c, c]$ to its entries

i.e. $B_{ij} = A_{ij} + z_{ij}$, $z_{ij} \sim [-10^{-9}, 10^{-9}]$

Poll what happens to the rank?

(a) rank $(B)=1$ (b) rank$(B)=2$ (c) rank$(B)=n$

unchanged doubles becomes max

Takeaway: when your matrix comes from data and/or subject to measurement error, its rank is large and not particularly useful

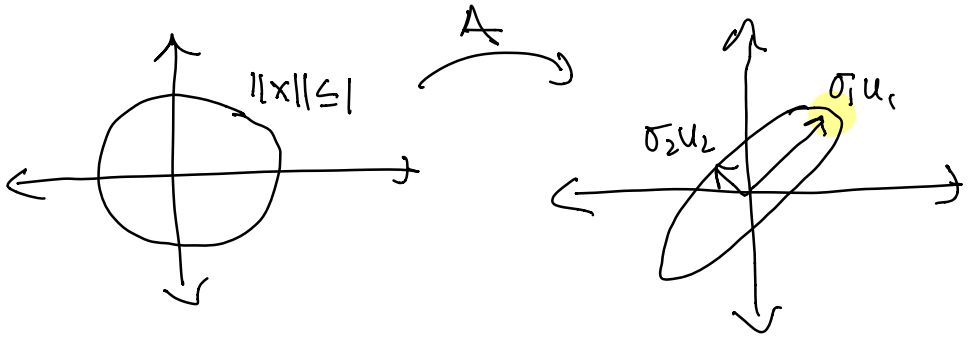
Great Idea: Even though rank is large, it is still close to a rank one matrix

First we need to define what "close" means

def: The operator norm of A, denote $\|A\|$, is $\max_{x \text{ s.t. } \|x\| \leq 1} \|Ax\|$

Note that when $\|\cdot\|$ is applied to a vector it is the length, but when it is applied to a matrix it is the operator norm

Recall the geometric view of SVD:



Q2: From this picture, what is the largest $\|Ax\|$ can be as we range over x with $\|x\| \leq 1$?

It is $\|\sigma_1 u_1\| = \sigma_1$, b/c u_i's are unit vectors and $\sigma_1 \geq \sigma_2 \ldots$

In fact we know even more from the SVD

Q3: What x achieves the max in the def. of the operator norm?

Let's try $x = v_1$. Then

$$Av_1 = \sum_{i=1}^{r} \sigma_i u_i v_i^T v_1$$

$$= \sigma_1 u_1$$

Fact: For any A, its operator norm is its largest singular value

There's another way to think about this:

Fact: For any matrix A and orthogonal matrix R we have $\|A\| = \|AR\|$

Why is this? You can check they are optimizing over the same set:

$$\max_{x \text{ s.t. } \|x\| \leq 1} \|Ax\| \quad \text{vs.} \quad \max_{z \text{ s.t. } \|z\| \leq 1} \|ARz\|$$

But if we set $x = Rz$ then $\|x\| = \|z\|$ and we can go back and forth

Also the same fact is true if we left multiply:

Fact: For any matrix A and orthogonal matrix R we have $\|A\| = \|RA\|$

So another way to think about why the SVD tells us the operator norm is:

$$\|A\| = \|U\Sigma V^T\| = \|\underbrace{U^T U}_{I}\Sigma V^T\|$$

$$= \|\Sigma V^T\| = \|\Sigma V^T V\| = \|\Sigma\|$$

Now since Σ is diagonal

$$\|\Sigma x\| = \sqrt{\sum_{i=1}^{r} \sigma_i^2 x_i^2}$$

and the largest you can make it is by setting $x_1 = 1, x_2 = 0, \ldots x_m = 0$

Now let's come back to our thought experiment

Q4: Maybe I can never recover A (rank one) from B exactly, but can I recover it approximately?

Since the noise we added was so small, we have:

$$\|B - A\| \text{ is very small}$$

So a natural strategy is to solve

$$\hat{A} = \underset{C \text{ s.t. } C \text{ is } n \times n \text{ and } \operatorname{rank}(C)=1}{\operatorname{argmin}} \|B - C\|$$

In words, we are looking for a rank one matrix that approximates B in the sense that the error matrix $B-C$ has small operator norm

Hope: We know that $C=A$ is a good solution so maybe the best C will be close to A

Before we get ahead of ourselves:

Q5: Can we solve this optimization problem?

we're searching over a complex, high-dimensional set — the set of all rank one matrices

Amazingly, the SVD contains the answer!

First consider the SVD of B

$$B = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

It is the sum of r rank 1 matrices

So let's take the largest one:

(in operator norm)

$$C = \sigma_1 u_1 v_1^T$$

Now let's compute how good this approx. is:

$$\|B-C\| = \|\sum_{i=2}^{r} \sigma_i u_i v_i^T\|$$

(*)

Q6: what is the operator norm of the error matrix?

well, the operator norm ≡ largest singular value, and our expression is an SVD for $B-C$, hence:

$$\|B-C\| = \sigma_2$$

Actually, this is the best you can do:

Theorem [Eckhart-Young] Let B be a matrix whose SVD is given by

$$B = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

Then $\min_{C \text{ s.t. } \operatorname{rank}(C) \le k} \|B-C\| = \sigma_{k+1}$ and it is achieved by $C = \sum_{i=1}^{k} \sigma_i u_i v_i^T$

This is called the truncated SVD

Last lecture we saw how the SVD tells you the rank

Key Takeaway: the SVD also tells you how close your matrix is to being low rank

Let's go back to our example and plot the singular values of B

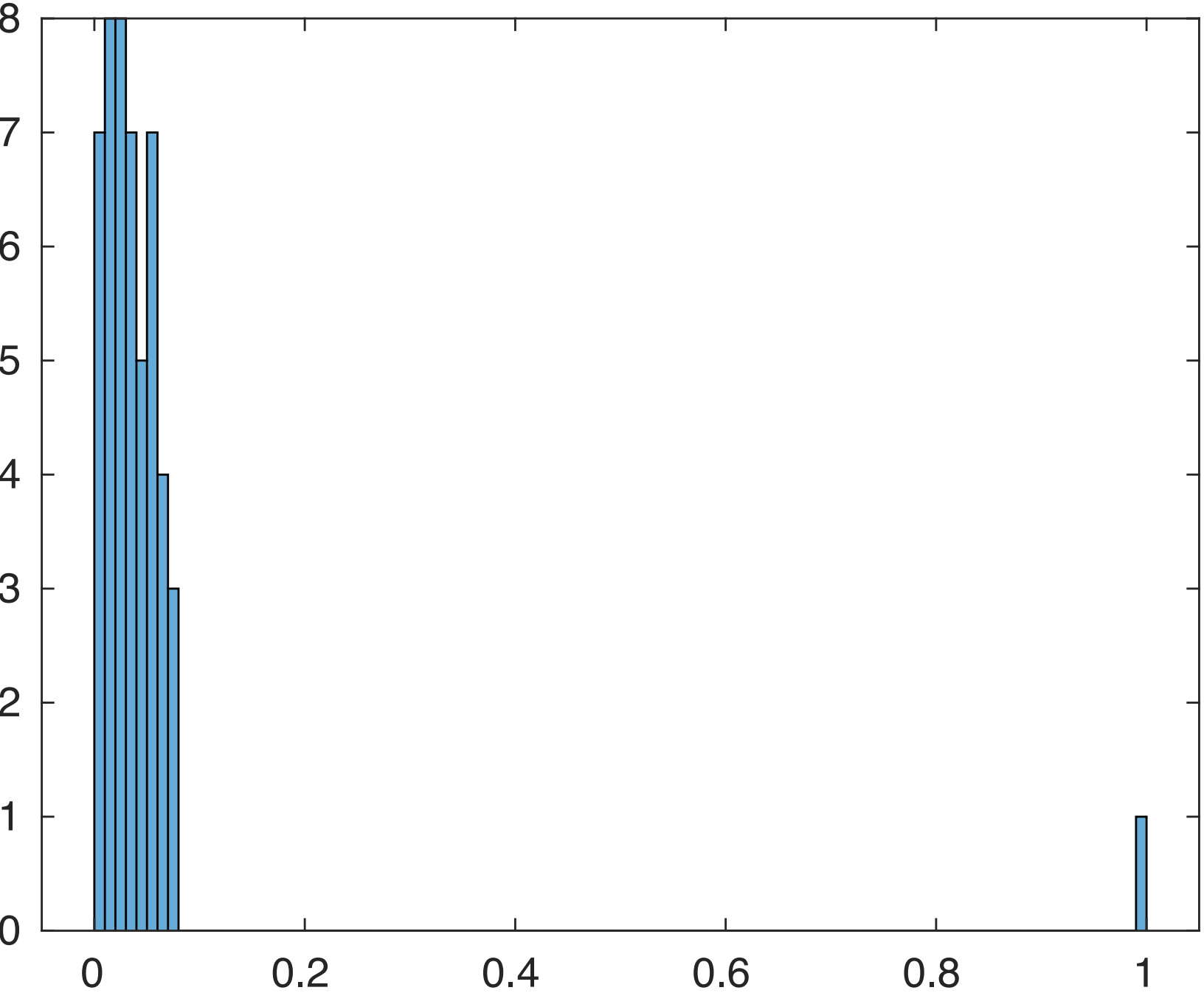
Experiment:

$$n = 50, \quad A = UV^T$$

(U, V: random $\pm\frac{1}{\sqrt{n}}$ entries)

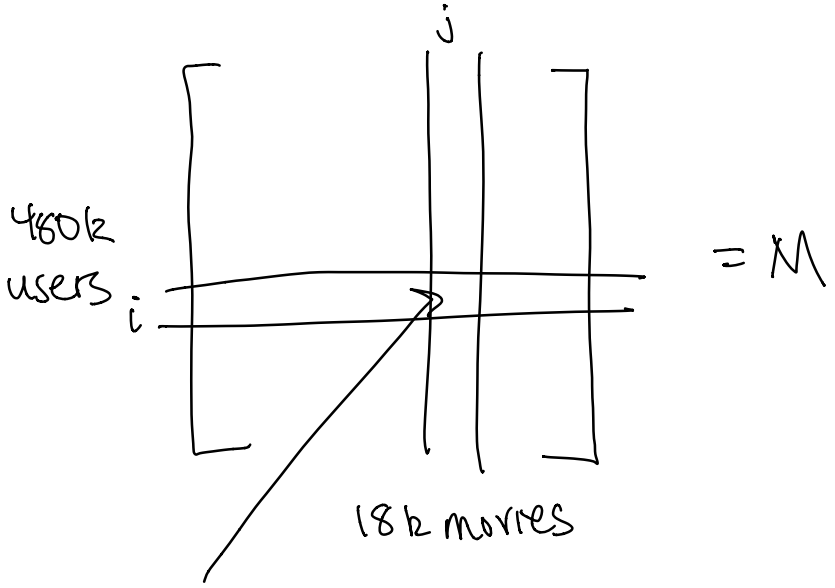
$$B_{ij} = A_{ij} + Z_{ij}, \quad Z_{ij} \sim_u [-10^{-2}, 10^{-2}]$$

"You can use the SVD to find structure in the presence of noisy / missing data"


8
7
6
5
4
3
2
1
0
0
0.2
0.4
0.6
0.8
1

Let's do one more example

Application: The Netflix Problem



How user i rated movie j (if at all), from 1-5 stars

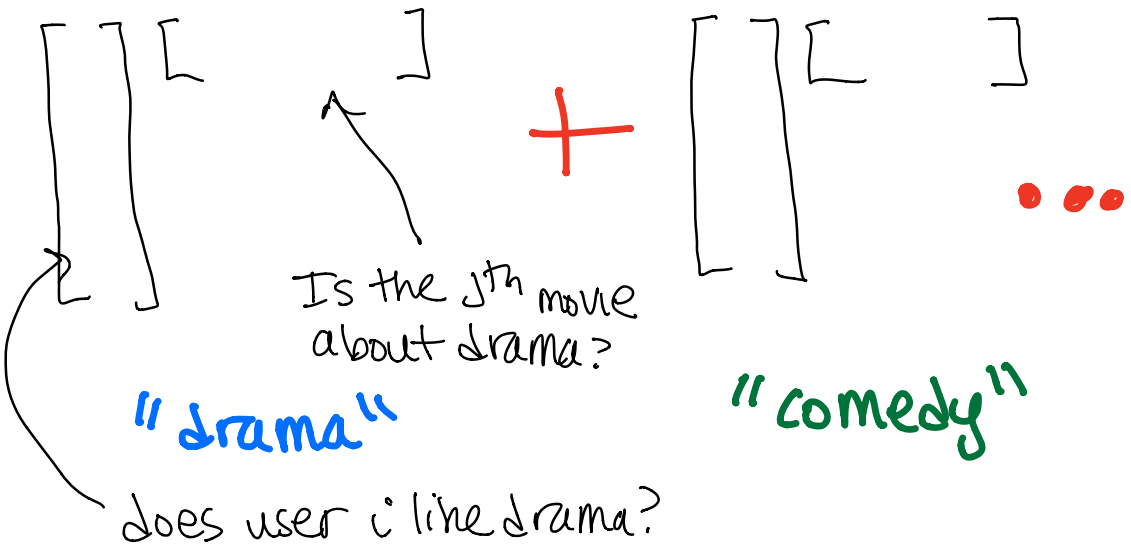
Problem: Only about 1% of the entries are observed

Q?: Can we estimate the missing entries?

This is called collaborative filtering

First, if we were to observe all of M would it have any useful structure?

Hypothesis: M is approximately low rank



Idea: Take the SVD of B where

$$B_{ij} = \begin{cases} M_{ij} & \text{for observed entries} \\ 0 & \text{else} \end{cases}$$

Then $B = M + \text{noise}$, and take the low rank approximation C and use C to predict missing entries/make recommendations