ELE 538B: Mathematics of High-Dimensional Data

Introduction

Yuxin Chen

Princeton University, Fall 2018

Big data

limited storage and processing ability

2.5 exabytes of data are generated every day (2012)

exabyte → zettabyte → yottabyte...??

Big data

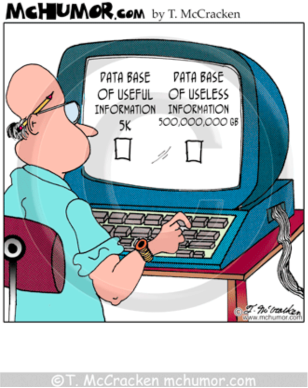
limited storage and processing ability

2.5 exabytes of data are generated every day (2012)

exabyte → zettabyte → yottabyte...??

Big data

limited storage and
processing ability



We're interested in the **information**
rather than the data

High-dimensional data analysis

Retrieve or infer information from high-dimensional / large-scale data

Challenges:

- High computational cost
- Only limited memory is available
- Do NOT want to compromise statistical accuracy

Low-dimensional structure

images

videos

U.S. COMMERCE'S ORTNER SAYS YEN UNDERVALUED

Commerce Dept. undersecretary of economic a®airs Robert Ortner said that he believed the dollar at current levels was fairly priced against most European currencies.

In a wide ranging address sponsored by the Export-Import Bank, Ortner, the bank's senior economist also said he believed that the yen was undervalued and could go up by 10 or 15 pct.

"I do not regard the dollar as undervalued at this point against the yen," he said.

On the other hand, Ortner said that he thought that "the yen is still a little bit undervalued," and "could go up another 10 or 15 pct."

In addition, Ortner, who said he was speaking personally, said he thought that the dollar against most European currencies was "fairly priced."

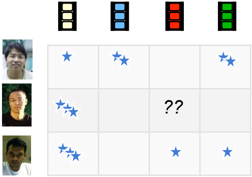
Ortner said his analysis of the various exchange rate values was based on such economic particulars as wage rate di®erentiations.

Ortner said there had been little impact on U.S. trade de¯cit by the decline of the dollar because at the time of the Plaza Accord, the dollar was extremely overvalued and that the ¯rst 15 pct decline had little impact.

He said there were indications now that the trade de¯cit was beginning to level o®.

Turning to Brazil and Mexico, Ortner made it clear that it would be almost impossible for those countries to earn enough foreign exchange to pay the service on their debts. He said the best way to deal with this was to use the policies outlined in Treasury Secretary James Baker's debt initiative.

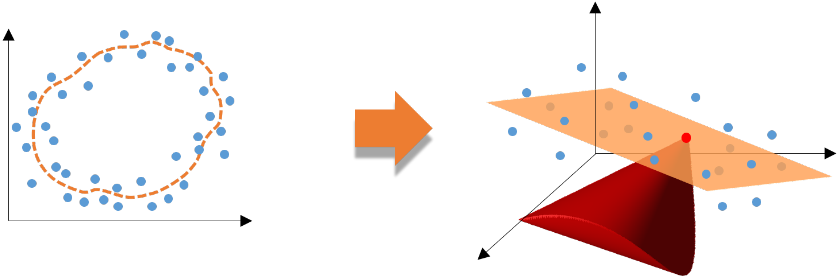
text



web data

Huge data sizes
but often
low-dimensional structure

Enabling efficient high-dimensional data analysis



Key: exploit low-dimensional geometry

Two examples of low-dimensional structure

- **Sparsity**

- **Low rank**

Sparsity

What is sparsity?

A signal is said to be sparse when most of its components vanish.

- Formally, $\boldsymbol{x} \in \mathbb{R}^p$ is said to be k-sparse if it has at most k nonzero entries

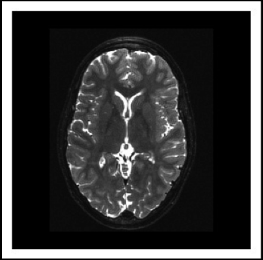
What is sparsity?

A signal is said to be sparse when most of its components vanish.

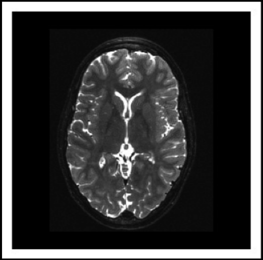
- Formally, $\boldsymbol{x} \in \mathbb{R}^p$ is said to be k-sparse if it has at most k nonzero entries
- Think of a k-sparse signal as having k degrees of freedom

Engineers wish to describe / approximate data in the most parsimonious terms!

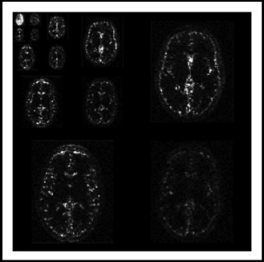
Only a small number of parameters matter



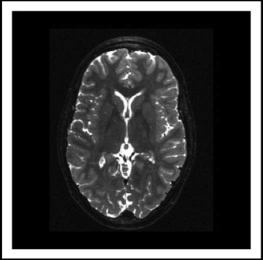
Only a small number of parameters matter



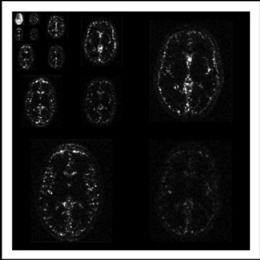
wavelet
trans-
form



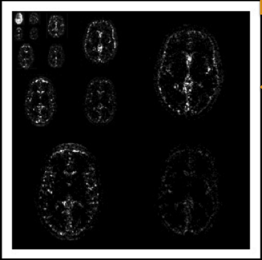
Only a small number of parameters matter



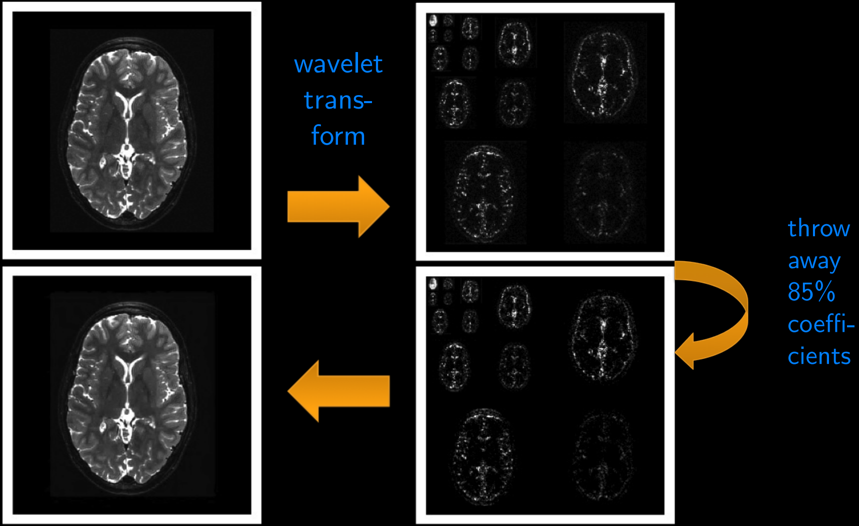
wavelet trans- form



throw away 85% coeffi- cients



Only a small number of parameters matter



Signal is very sparse in some transform domain (e.g. wavelet)

Only a small number of parameters matter

- Compute 10^6 wavelet coeffients
- Keep only the $25K$ largest coefficients
- Inverse wavelet transform

1 megapixel image

25k term approximation

Only a small number of parameters matter

Raw: 15MB

Only a small number of parameters matter

Raw: 15MB

JPEG: 150KB

There is (almost) no loss in quality between the raw image and its JPEG compressed form

General philosophy

We are drowning in information and starving for knowledge

Rutherford Roger

- Massive data acquisition
- Most data is redundant and can be thrown away

General philosophy

We are drowning in data and starving for information

Rutherford Roger

- Massive data acquisition
- Most data is redundant and can be thrown away

General philosophy

We are drowning in data and starving for information

Rutherford Roger

- Massive data acquisition
- Most data is redundant and can be thrown away

Will such "information sparsity" be useful in data acquisition, statistical inference and information recovery?

Advantages of sparsity

- Interpretation of our estimate / fitted model
  - particularly important when sample size $\ll$ $\#$ unknowns

Advantages of sparsity

- Interpretation of our estimate / fitted model
  - particularly important when sample size $\ll \#$ unknowns
- Computational convenience
  - in many cases we have scalable procedures to promote sparsity

Advantages of sparsity

- Interpretation of our estimate / fitted model
  - particularly important when sample size $\ll$ # unknowns
- Computational convenience
  - in many cases we have scalable procedures to promote sparsity
- "Bet on sparsity" principle
  - *use a procedure that does well in sparse problems, since no procedure does well in dense problems*
  - "less is more": sparse model might be easier to estimate than dense models

Example: compressed sensing

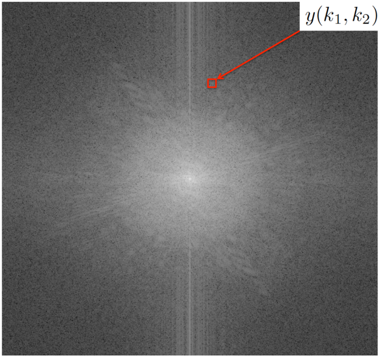
Magnetic Resonance Imaging (MRI)

MR scanner

MR image

K. Pauly, G. Gold, RAD220

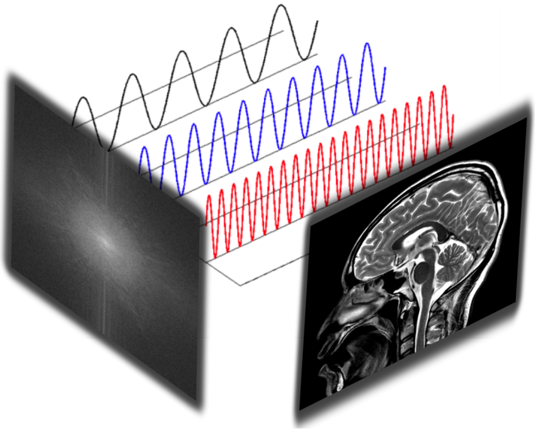
What an MRI machine sees



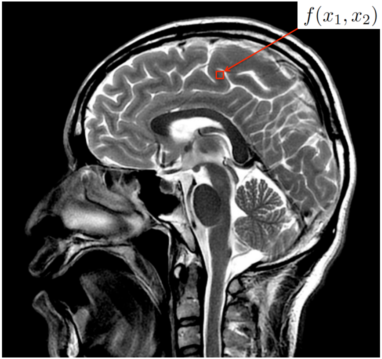
Measured data $y(k_1, k_2)$ $\longleftarrow$ Fourier transform of image $f(x_1, x_2)$

Fourier transform

$$y(k_1, k_2) \approx \sum_{x_1} \sum_{x_2} f(x_1, x_2) \mathrm{e}^{-i2\pi(k_1 x_1 + k_2 x_2)}$$



How do we form an image?



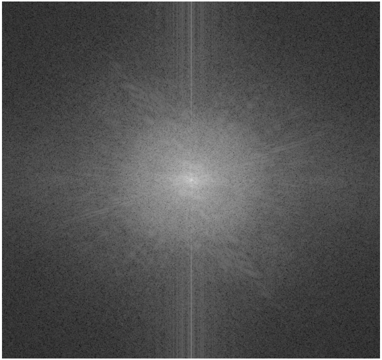
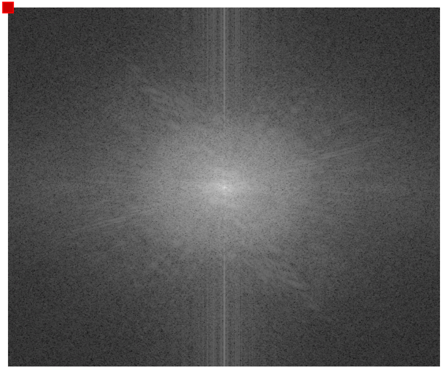


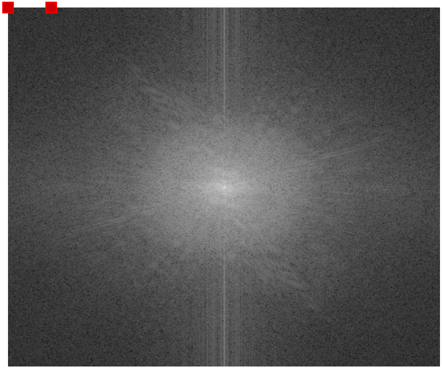
image $f(x_1, x_2)$ $\longleftarrow$ inverse Fourier transform of measurements

$$f(x_1, x_2) \approx \sum \sum y(k_1, k_2) \mathrm{e}^{i2\pi(k_1 x_1 + k_2 x_2)}$$

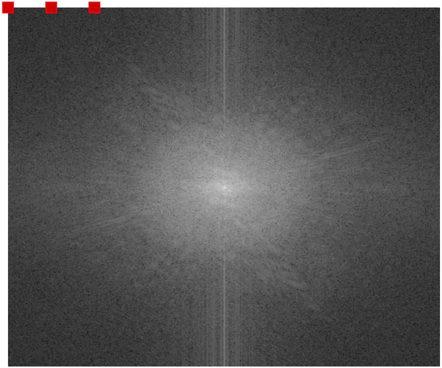
MRI data collection is inherently slow



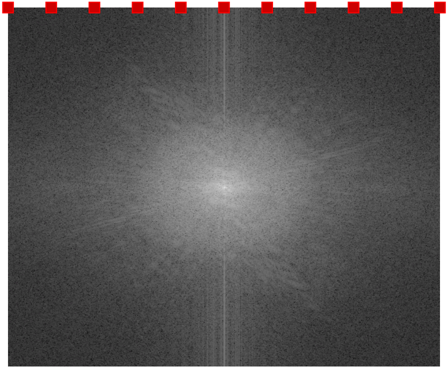
MRI data collection is inherently slow



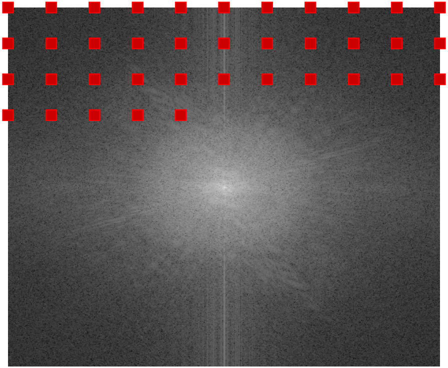
MRI data collection is inherently slow



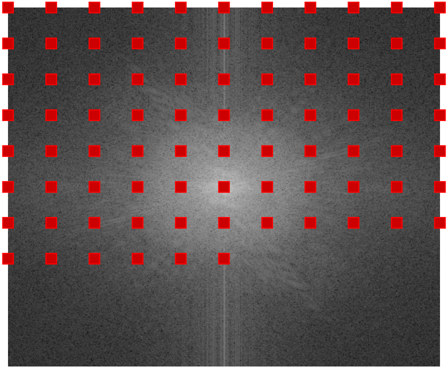
MRI data collection is inherently slow



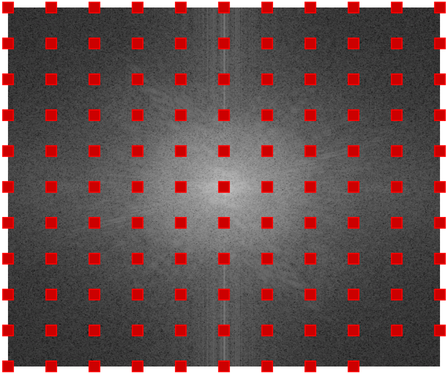
MRI data collection is inherently slow



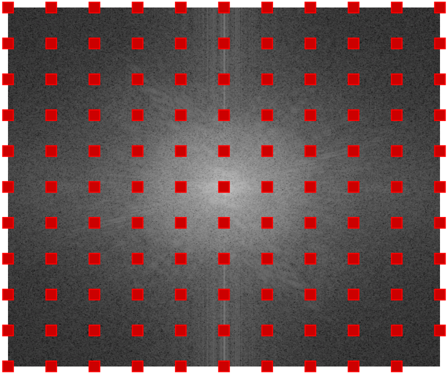
MRI data collection is inherently slow



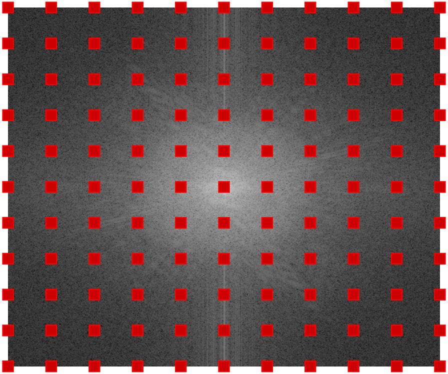
MRI data collection is inherently slow



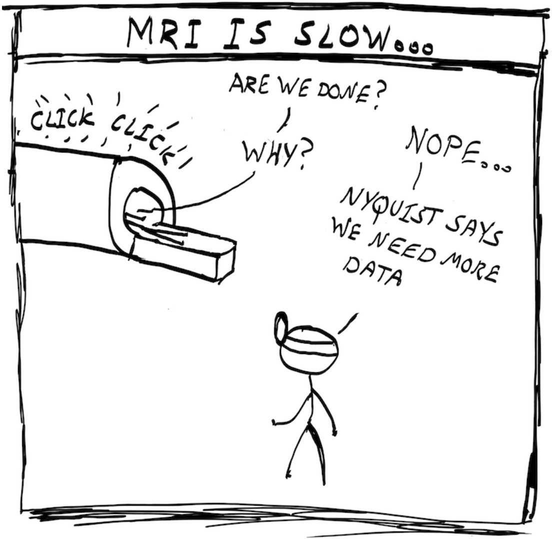
MRI data collection is inherently slow



MRI data collection is inherently slow



Done!


MRI IS SLOW...
CLICK CLICK CLICK
ARE WE DONE?
WHY?
NOPE...
NYQUIST SAYS WE NEED MORE DATA

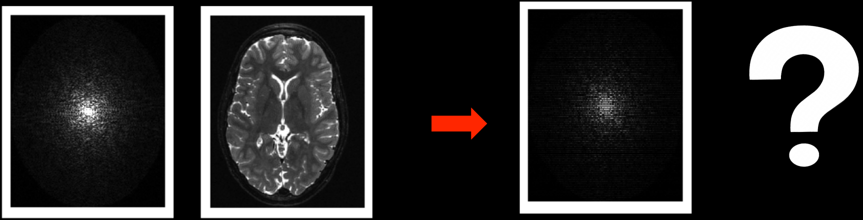
M. Lustig

Fact: impact of MRI on children health is limited

Children cannot stay still or breathhold!

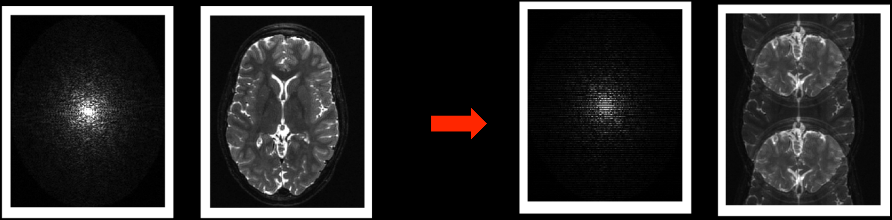
- (deep) anesthesia required
- respiration suspension

Is it possible to take fewer samples to reduce scan time?



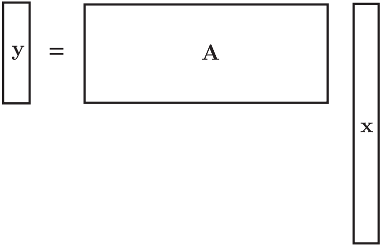
uniform undersampling by a factor of 2

Is it possible to take fewer samples to reduce scan time?



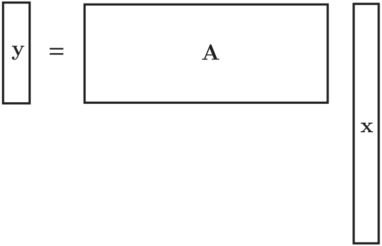
uniform undersampling by a factor of 2

Fewer equations than unknowns!



How can we possibly solve an underdetermined system?

Fewer equations than unknowns!



How can we possibly solve an underdetermined system?

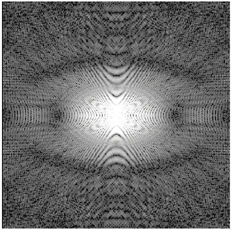
We need at least as many equations as unknowns!

Carl Friedrich Gauss

A surprising experiment

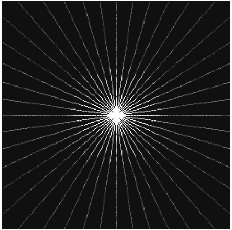
A surprising experiment

Fourier transform



A surprising experiment

Fourier transform

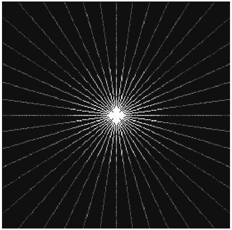


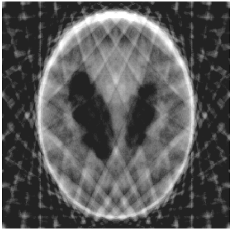
highly subsampled

A surprising experiment

classical
reconstruction

Fourier transform

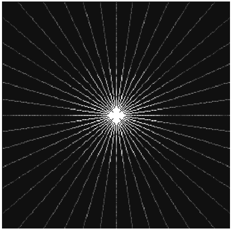




highly subsampled

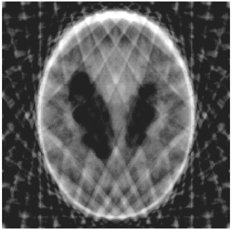
A surprising experiment

Fourier transform



highly subsampled

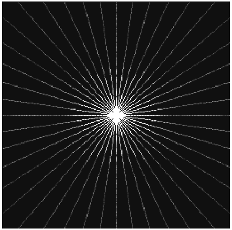
classical
reconstruction



compressed sensing
reconstruction

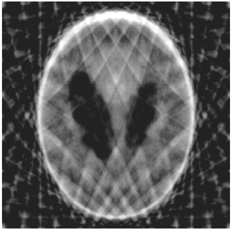
A surprising experiment

Fourier transform



highly subsampled

classical reconstruction

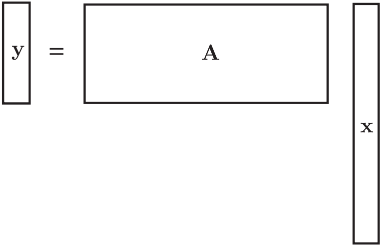


compressed sensing reconstruction

CS algorithm:

$$\min \quad \sum_x ||\nabla f(x)||_1 \text{ subj. to data constraints}$$

Structured solutions

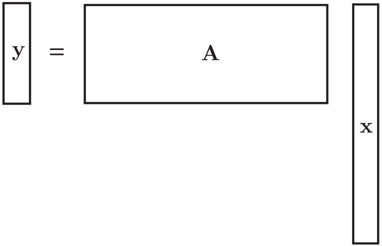


How can we possibly solve?

Need some structure

$\boldsymbol{x}$ is k-sparse $\rightarrow$ at most k degrees of freedom

Ingredients for success



- Exploit signal structure: sparsity
- Recovery via efficient algorithms (e.g. convex optimization)
- Incoherent sensing mechanism

Low-rank structure

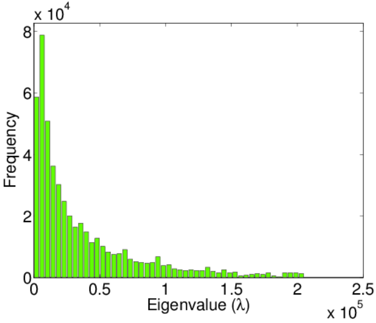
Netflix challenge: predict unseen ratings

Can we infer the missing entries?

$$\begin{bmatrix} \checkmark & ? & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \\ \checkmark & ? & ? & \checkmark & ? & ? \\ ? & ? & \checkmark & ? & ? & \checkmark \\ \checkmark & ? & ? & ? & ? & ? \\ ? & \checkmark & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \end{bmatrix}$$

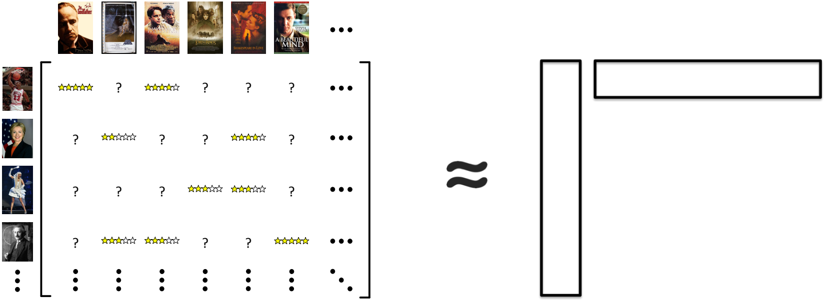
- Underdetermined system (more unknowns than revealed entries)
- Seems hopeless

What if unknown matrix has structure?



A few factors explain most of the data

What if unknown matrix has structure?

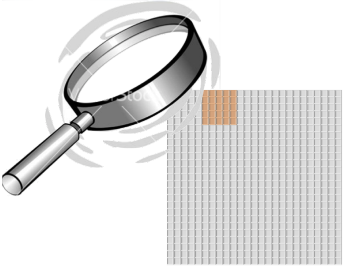


A few factors explain most of the data $\longrightarrow$ low-rank approximation

Low-rank matrix completion?

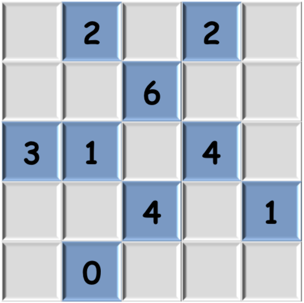
3	2	4	2	1
4	2	6	4	2
3	1	5	4	2
3	1	4	3	1
1	0	3	3	2

Ground truth



50×50 low-rank matrix

Another surprising experiment



Observed samples

Another surprising experiment

	2		2	
		6		
3	1		4	
		4		1
	0			

Observed samples

3	2	4	2	1
4	2	6	4	2
3	1	5	4	2
3	1	4	3	1
1	0	3	3	2

Estimate via nuclear norm min

minimize $\underbrace{\text{sum-of-singular-values}}_{\text{nuclear norm}}$ subj. to data constraints

Another surprising experiment

3	2	4	2	1
4	2	6	4	2
3	1	5	4	2
3	1	4	3	1
1	0	3	3	2

Ground truth

3	2	4	2	1
4	2	6	4	2
3	1	5	4	2
3	1	4	3	1
1	0	3	3	2

Estimate via nuclear norm min

minimize $\underbrace{\text{sum-of-singular-values}}_{\text{nuclear norm}}$ subj. to data constraints

Main theme of this course

- Low-dimensional structure (e.g. sparsity, low rank)

Main theme of this course

- Low-dimensional structure (e.g. sparsity, low rank)
- Incoherent sensing mechanism (often designed via "random" sampling)

Main theme of this course

- Low-dimensional structure (e.g. sparsity, low rank)
- Incoherent sensing mechanism (often designed via "random" sampling)
- Efficient algorithms

Main theme of this course

- Low-dimensional structure (e.g. sparsity, low rank)
- Incoherent sensing mechanism (often designed via “random” sampling)
- Efficient algorithms

We can estimate many low-dimensional structures of interest from highly incomplete data by efficient algorithms

Logistics

Why you should not take this course

Why you **should not** take this course

- There will be quite a few THEOREMS and PROOFS ...

Why you **should not** take this course

- There will be quite a few THEOREMS and PROOFS ...

- Nonrigorous / heuristic from time to time

Why you **should** consider taking this course

Why you should consider taking this course

- There will be quite a few THEOREMS and PROOFS ...

Why you should consider taking this course

- There will be quite a few THEOREMS and PROOFS ...
  - promote deeper understanding of scientific results

Why you **should** consider taking this course

- There will be quite a few THEOREMS and PROOFS ...
    - promote deeper understanding of scientific results
- Nonrigorous / heuristic from time to time

Why you should consider taking this course

- There will be quite a few THEOREMS and PROOFS ...
  - promote deeper understanding of scientific results
- Nonrigorous / heuristic from time to time
  - "nonrigorous" but grounded in rigorous theory
  - help develop intuition

Tentative topics

- Spectral methods
- Matrix concentration inequalities
- Large-scale numerical linear algebra
- Tensor decomposition
- Sparse recovery and compressed sensing
- Low-rank matrix recovery and robust PCA
- Nonconvex matrix factorization
- Super-resolution and spectral estimation

Tentative topics

- Spectral methods
- Matrix concentration inequalities
- Large-scale numerical linear algebra
- Tensor decomposition
- Sparse recovery and compressed sensing
- Low-rank matrix recovery and robust PCA
- Nonconvex matrix factorization
- Super-resolution and spectral estimation
- *Shallow neural networks?*

Topics that I won't have time to cover

- Sparse representation
- Model selection
- Large-scale optimization methods (my other course)

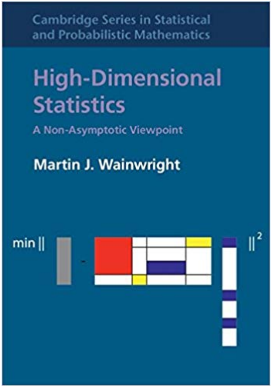
Topics that I won't have time to cover

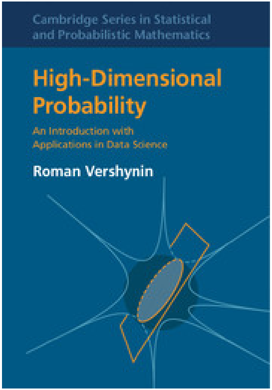
- Sparse representation
- Model selection
- Large-scale optimization methods (my other course)

Slides will be provided if you are interested ...

Textbooks

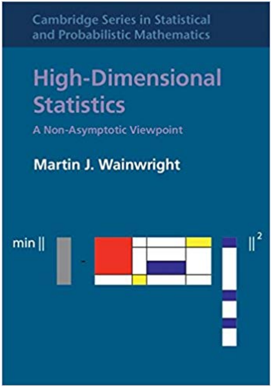
We recommend these books, but will not follow them closely ...



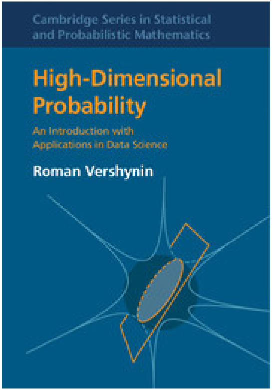


Textbooks

We recommend these books, but will not follow them closely ...



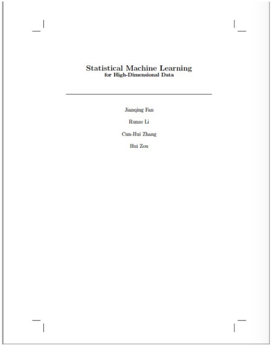
to be published next year ...

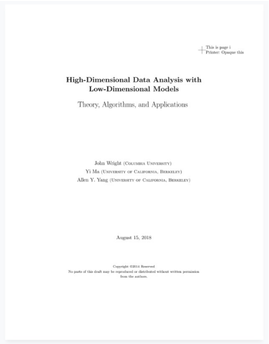


to be published this year ...

Textbooks

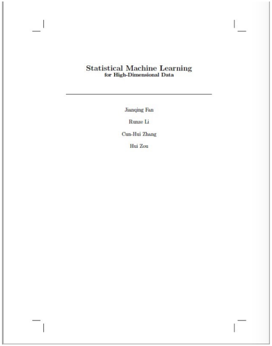
We recommend these books, but will not follow them closely ...





Textbooks

We recommend these books, but will not follow them closely ...


Statistical Machine Learning
for High-Dimensional Data

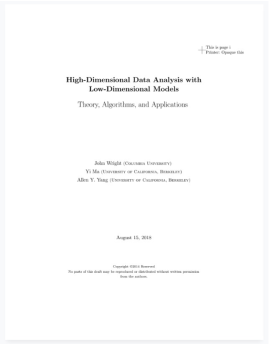
Jianqing Fan

Runze Li

Cun-Hui Zhang

Hui Zou

to be published in ??? ...


This is page i
Printer: Opaque this

High-Dimensional Data Analysis with
Low-Dimensional Models

Theory, Algorithms, and Applications

John Wright (Columbia University)
Yi Ma (University of California, Berkeley)
Allen Y. Yang (University of California, Berkeley)

August 15, 2018

Copyright ©2014 Reserved
No parts of this draft may be reproduced or distributed without written permission
from the authors.

to be published in ??? ...

Prerequisites

- basic linear algebra
- basic probability
- a programming language (e.g. Matlab, Python, ...)
- *knowledge in basic convex optimization*

Grading

- Homeworks (40%): ∼3 problem sets
  - Use Piazza as the main mode of electronic communication; please post (and answer) questions there!
- Term project (60%)
  - Either individually or in groups of two

Term project

Two forms

- literature review
- original research
  - *You are strongly encouraged to combine it with your own research*

Term project

Two forms

- literature review
- original research
  - *You are strongly encouraged to combine it with your own research*

Three milestones

- Proposal (Oct 19): up to 1 page
- Presentation (last week of class)
- Report (Jan 14): up to 4 pages with unlimited appendix

Asymptotic notation used in this course

- $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ means

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} \;\leq\; \text{const}$$

- $f(n) \gtrsim g(n)$ or $f(n) = \Omega(g(n))$ means

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} \;\geq\; \text{const}$$

- $f(n) \asymp g(n)$ or $f(n) = \Theta(g(n))$ means

$$\text{const}_1 \;\leq\; \lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} \;\leq\; \text{const}_2$$

- $f(n) = o(g(n))$ means

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} \;=\; 0$$

Reference

[1] "*Mathematics of sparsity (and a few other things)*," E. Candes, *International Congress of Mathematicians*, 2014.

[2] "*Statistical learning with sparsity: the Lasso and generalizations*," T. Hastie, R. Tibshirani, and M. Wainwright, 2015.