Sparse and Low-Rank Representation Lecture I: Motivation and Theory

Yi Ma

MSRA and UIUC

Allen Yang

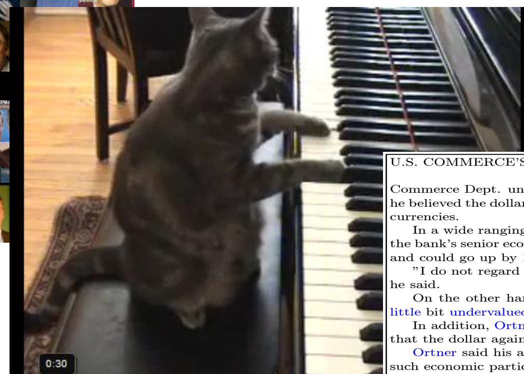
UC Berkeley

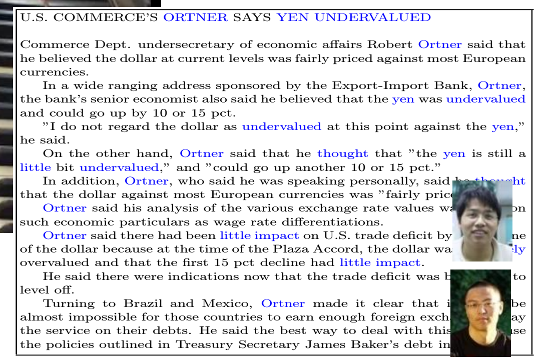
John Wright

Columbia University

European Conference on Computer Vision, October 7, 2012

CONTEXT – *Data increasingly massive, high-dimensional…*





U.S. COMMERCE'S ORTNER SAYS YEN UNDERVALUED

Commerce Dept. undersecretary of economic affairs 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."

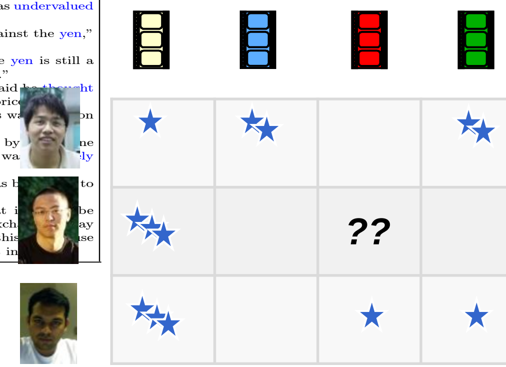
In addition, Ortner, who said he was speaking personally, said he thought that the dollar against most European currencies was "fairly pric…

Ortner said his analysis of the various exchange rate values wa… …on such economic particulars as wage rate differentiations.

Ortner said there had been little impact on U.S. trade deficit by … …ne of the dollar because at the time of the Plaza Accord, the dollar wa… …ly overvalued and that the first 15 pct decline had little impact.

He said there were indications now that the trade deficit was b… …to level off.

Turning to Brazil and Mexico, Ortner made it clear that i… …be almost impossible for those countries to earn enough foreign exch… …ay the service on their debts. He said the best way to deal with this… …se the policies outlined in Treasury Secretary James Baker's debt in…



Images

- ➢ ***1M pixels***

Videos

- ➢ ***1B voxels***

Web data

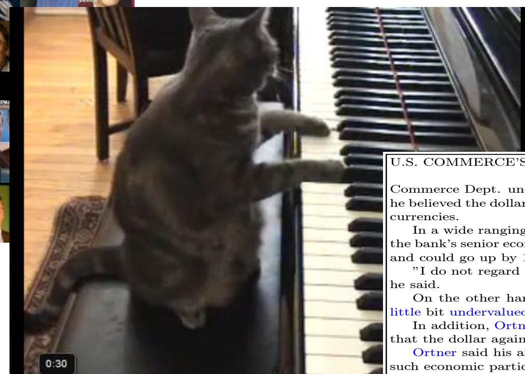
- ➢ ***100B webpages***

User data

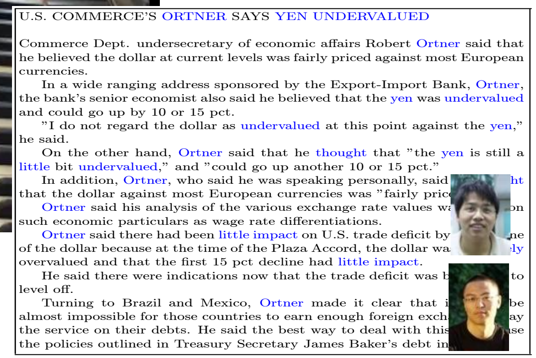
- ➢ ***1B users***

CONTEXT – *Discovering knowledge from data*

Images



Videos



U.S. COMMERCE'S ORTNER SAYS YEN UNDERVALUED

Commerce Dept. undersecretary of economic affairs 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 [illegible] that the dollar against most European currencies was "fairly pric[illegible]

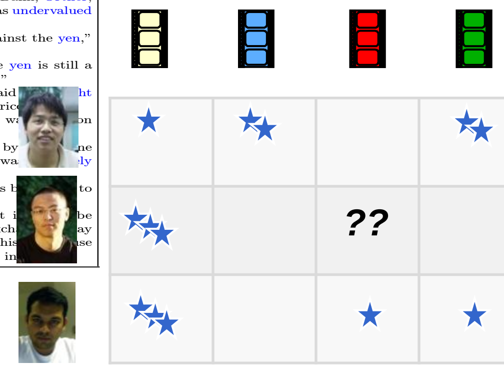
Ortner said his analysis of the various exchange rate values w[illegible] such economic particulars as wage rate differentiations.

Ortner said there had been little impact on U.S. trade deficit by [illegible] of the dollar because at the time of the Plaza Accord, the dollar wa[illegible] overvalued and that the first 15 pct decline had little impact.

He said there were indications now that the trade deficit was b[illegible] level off.

Turning to Brazil and Mexico, Ortner made it clear that i[illegible] almost impossible for those countries to earn enough foreign exch[illegible] the service on their debts. He said the best way to deal with this [illegible] the policies outlined in Treasury Secretary James Baker's debt in[illegible]

Web data



User data

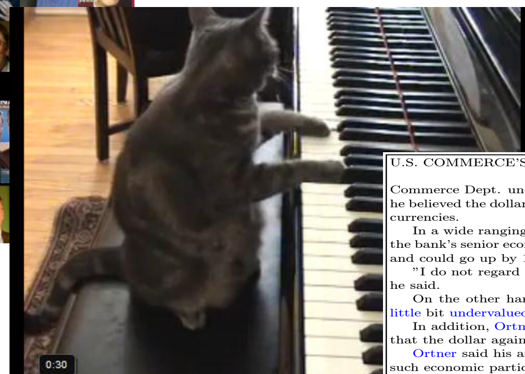
How to extract compact knowledge from such massive datasets?

CONTEXT – *Good solutions impact many applications*

Images

⇓

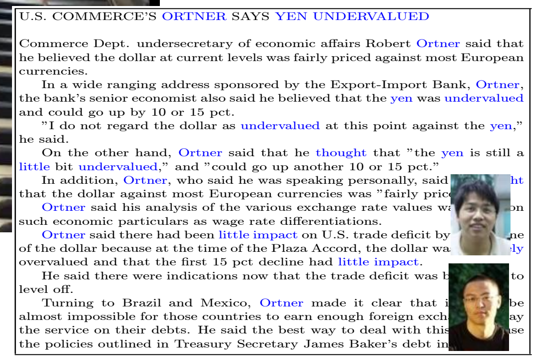
Compression
Denoising
Superresolution
Recognition…



Videos

⇓

Streaming
Tracking
Stabilization…



U.S. COMMERCE'S ORTNER SAYS YEN UNDERVALUED

Commerce Dept. undersecretary of economic affairs 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 ht that the dollar against most European currencies was "fairly pric

Ortner said his analysis of the various exchange rate values wa on such economic particulars as wage rate differentiations.

Ortner said there had been little impact on U.S. trade deficit by ne of the dollar because at the time of the Plaza Accord, the dollar wa ly overvalued and that the first 15 pct decline had little impact.

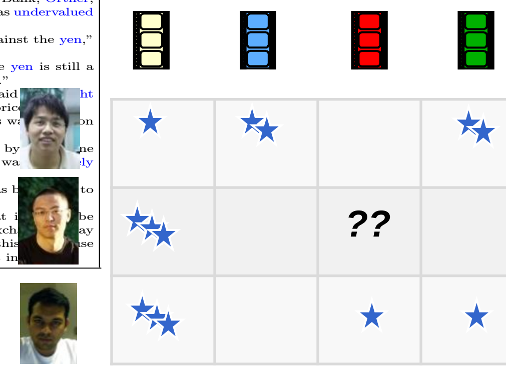
He said there were indications now that the trade deficit was b to level off.

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

Web data

⇓

Indexing
Ranking
Search…

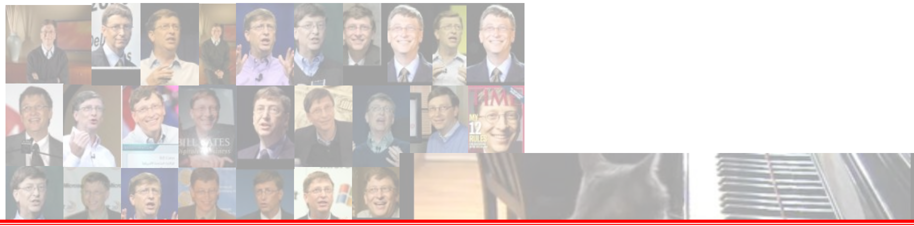


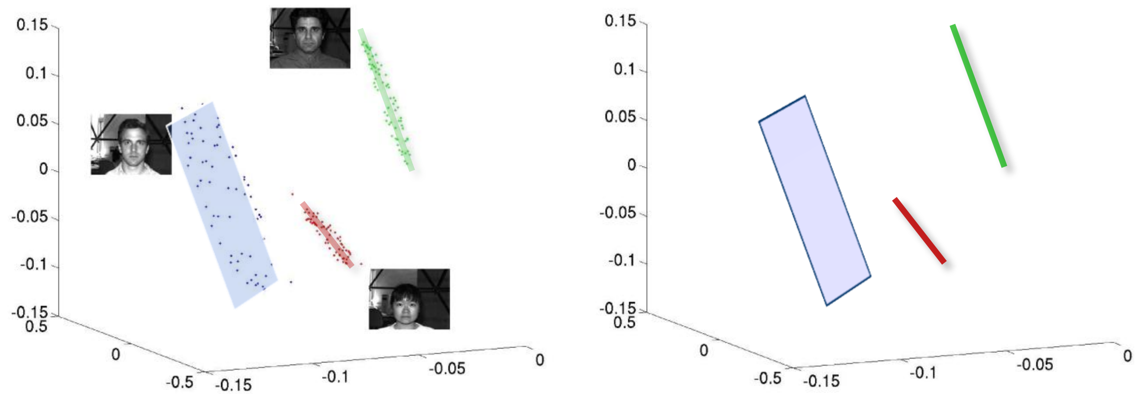
User data

⇓

Clustering
Classification
Collaborative filtering…

Low-dimensional structures in high-dimensional data

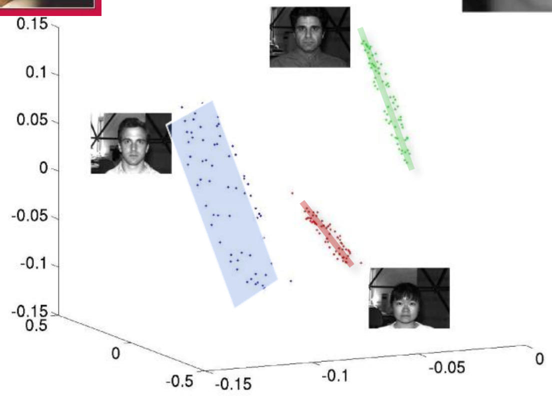


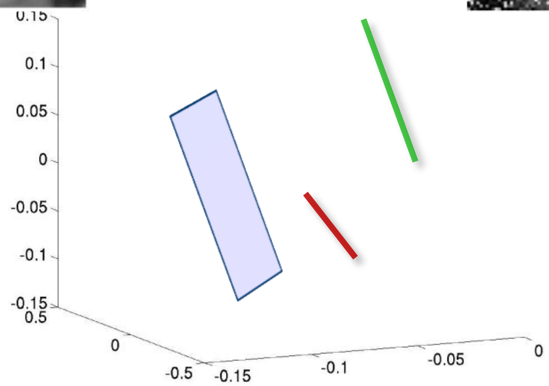


How can we learn and exploit low-dimensional structures in high-dimensional data?

But it is not so easy…

?





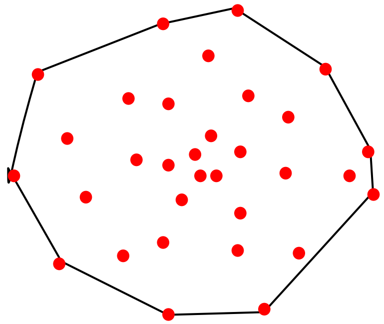
Real application data often contain missing observations, corruptions, or even malicious errors.

Classical methods (e.g., least squares, PCA) break down…

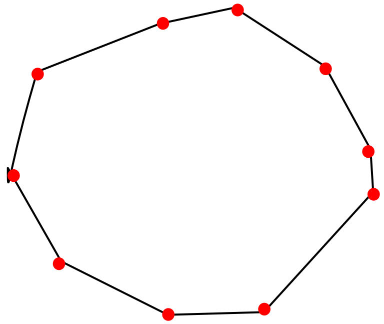
CONTEXT - *New Phenomena with High-Dimensional Data*

KEY CHALLENGE: efficiently and reliably recover low-dimensional structures from high-dimensional data, despite gross observation errors.

A sobering message: human intuition is severely limited in high-dimensional spaces:



Gaussian samples in 2D

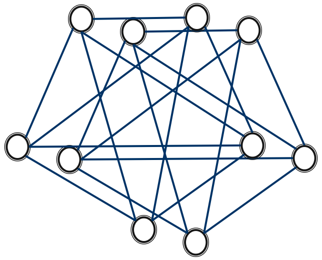
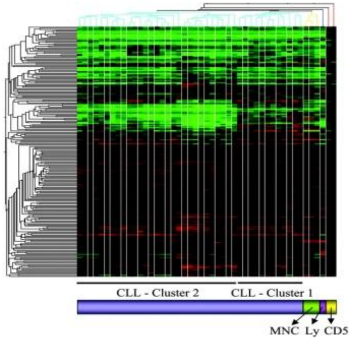


As dimension grows proportionally with the number of samples…

A **new regime** of geometry, statistics, and computation…

CONTEXT - *Massive High-Dimensional Data*



Recognition **Surveillance** **Search and Ranking** **Bioinformatics**

The curse of dimensionality:
…increasingly demand inference with limited samples for very high-dimensional data.

The blessing of dimensionality:
… real data highly concentrate on low-dimensional, sparse, or degenerate structures in the high-dimensional space.

But **nothing is free**: *Gross errors and irrelevant measurements are now ubiquitous in massive cheap data.*

A long and rich history of robust estimation with error correction and missing data imputation:

R. J. Boscovich. ***De calculo probailitatum que respondent diversis valoribus summe errorum post plures observationes … , before 1756***

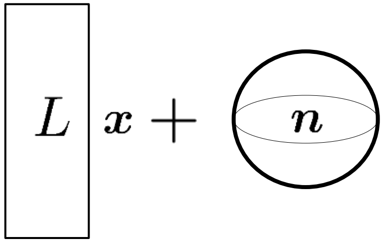
A. Legendre. ***Nouvelles methodes pour la determination des orbites des cometes*, 1806**

C. Gauss. ***Theory of motion of heavenly bodies, 1809***

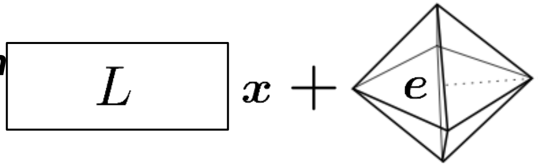
A. Beurling. ***Sur les integrales de Fourier absolument convergentes et leur application a une transformation functionelle*, 1938**

B. Logan. ***Properties of High-Pass Signals*, 1965**

⋮



over-determined
+ dense, Gaussian

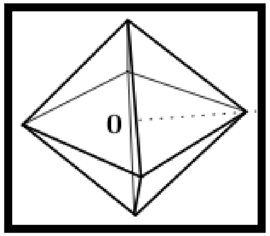


underdetermined
+ sparse, Laplacian

… IS NEW AGAIN

Today, robust estimation in high dimensions is more urgent, more tractable, and increasingly sharply understood.

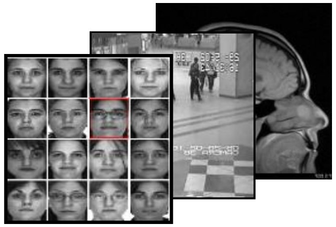
Theory – high-dimensional geometry & statistics, measure concentration, combinatorics, coding theory…



Algorithms – large scale convex optimization, parallel and distributed computing….

$$\min \|\boldsymbol{x}\|_\diamond \text{ s.t. } \boldsymbol{y} = A\boldsymbol{x}$$

Applications – massive data driven methods, sensing and hashing, denoising, superresolution, MRI, bioinformatics, image classification, recognition …



Today's plan

Lecture I: Motivation and Theory

Lecture II: Efficient Optimization for Sparse Representation

Lecture III: Applications and Generalizations

Q&A and discussion

Lecture I

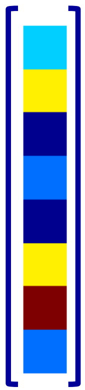
Theory of Sparse and Low-Rank Recovery

John Wright

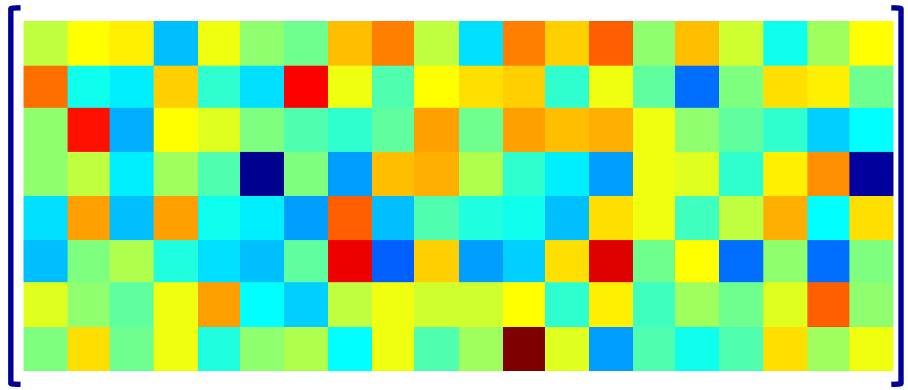
Electrical Engineering

Columbia University

UNDERDETERMINED LINEAR SYSTEMS



$=$



$$\begin{bmatrix} ? \\ ? \\ \vdots \\ ? \\ ? \end{bmatrix}$$

Observation

$\boldsymbol{y} \in \mathbb{R}^m$

$\boldsymbol{A} \in \mathbb{R}^{m \times n}$

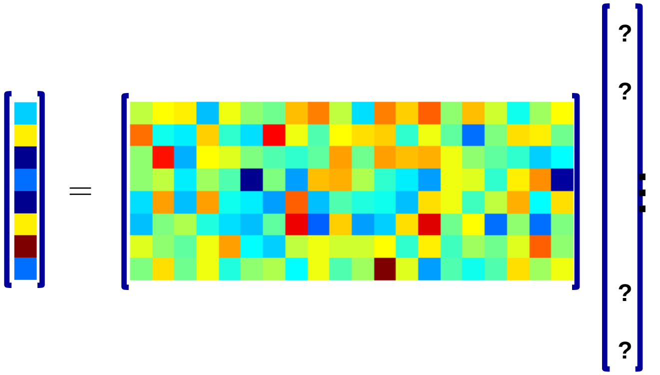
Unknown

$\boldsymbol{x} \in \mathbb{R}^n$

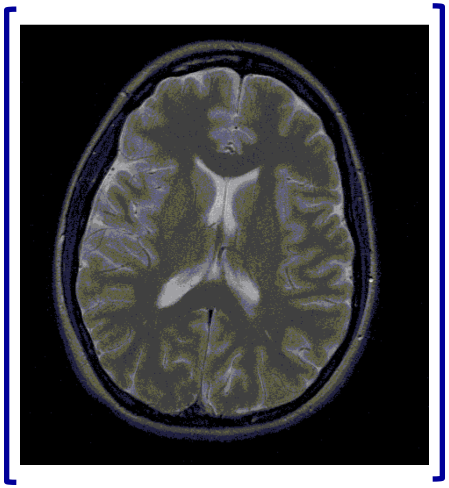
UNDERDETERMINED LINEAR SYSTEMS

Underdetermined system

$$y = Ax$$



Signal acquisition



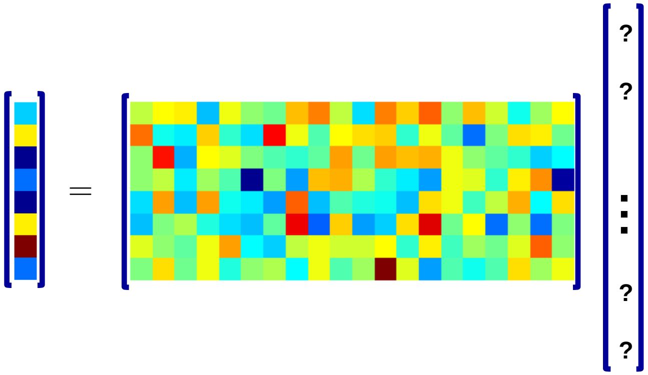
z

Image to be sensed

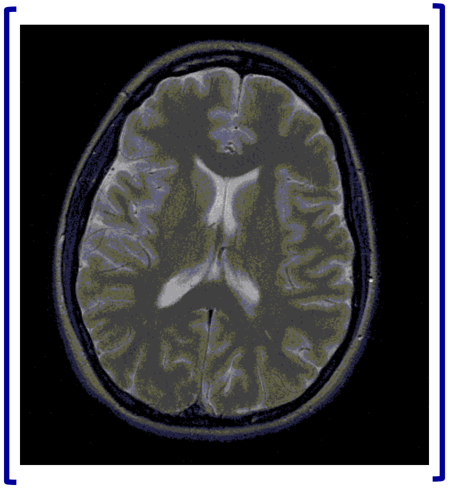
UNDERDETERMINED LINEAR SYSTEMS

Underdetermined system

$$y = Ax$$



Signal acquisition



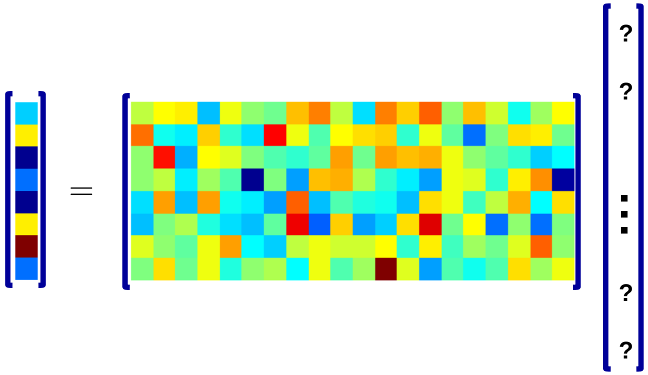
z

Image to be sensed

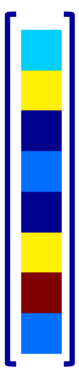
UNDERDETERMINED LINEAR SYSTEMS

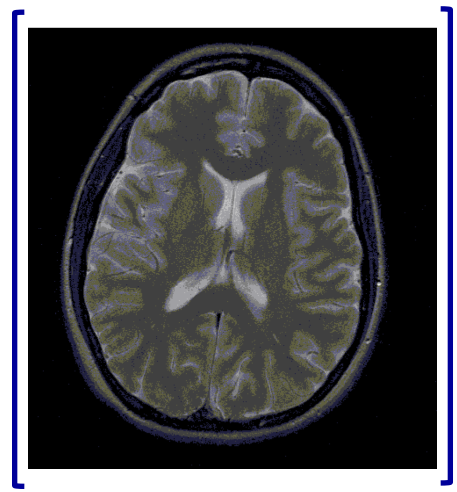
Underdetermined system

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$



Signal acquisition

$\boldsymbol{y}$  $=$ 



$\boldsymbol{z}$

$$y_i = \int_{\boldsymbol{u}} \boldsymbol{z}(u) \exp(-2\pi j \boldsymbol{k}(t_i)^* \boldsymbol{u}) d\boldsymbol{u}$$

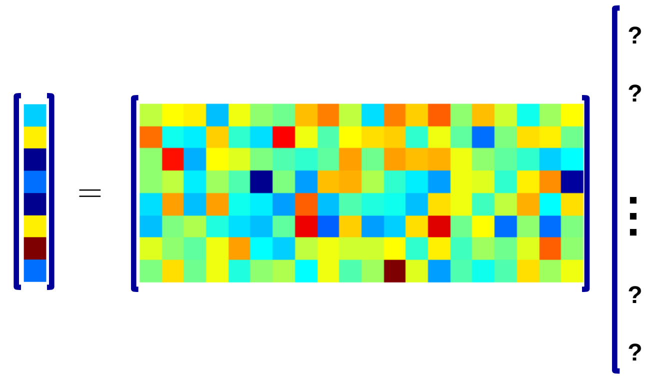
Observations are Fourier coefficients!

Image to be sensed

UNDERDETERMINED LINEAR SYSTEMS

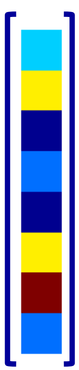
Underdetermined system

$$y = Ax$$



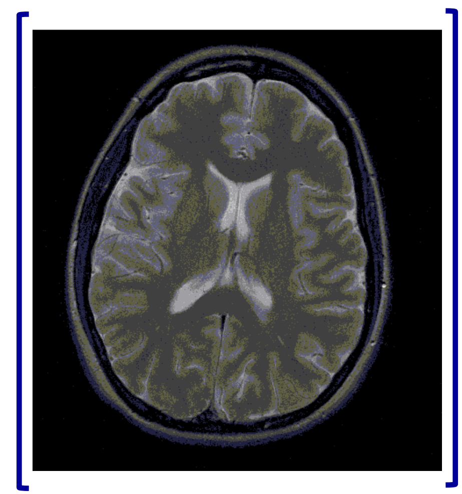
Signal acquisition

y



$=$

F_Ω



z

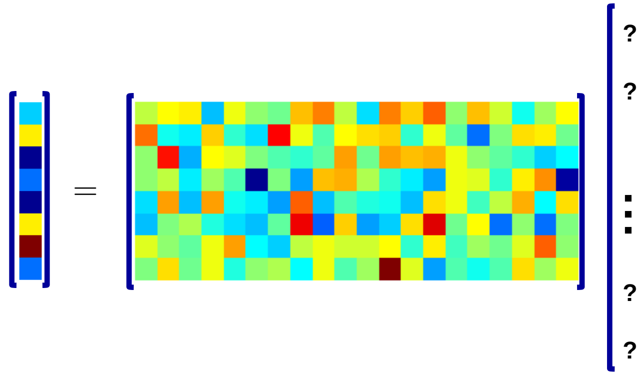
A few Fourier coefficients

Image to be sensed

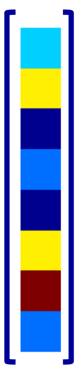
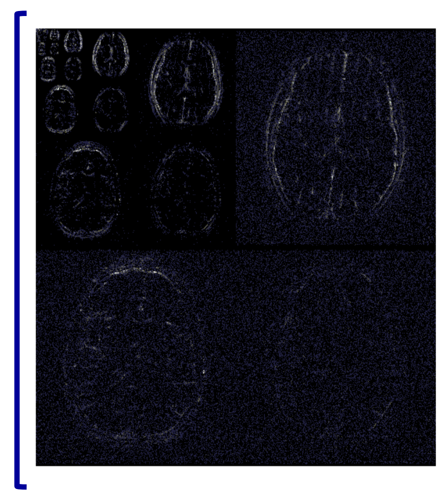
UNDERDETERMINED LINEAR SYSTEMS

Underdetermined system

$$y = Ax$$



Signal acquisition

y  $=$  Ψ 

F_Ω

x

A few Fourier coefficients

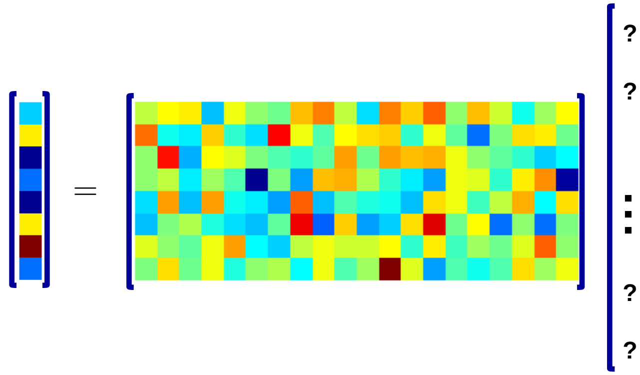
Wavelet coefficients: $z = \Psi x$

[Lustig, Donoho + Pauly '10] … brain image – Lustig '12

UNDERDETERMINED LINEAR SYSTEMS

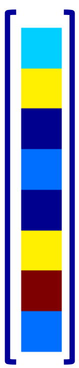
Underdetermined system

$$y = Ax$$

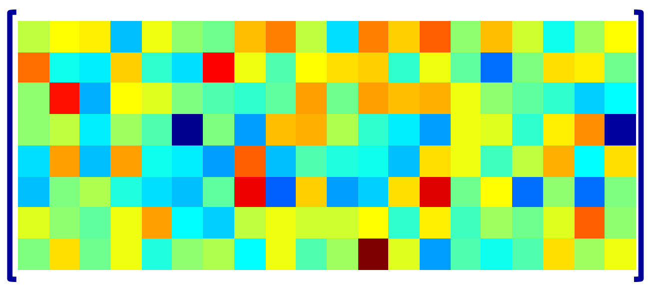


Signal acquisition

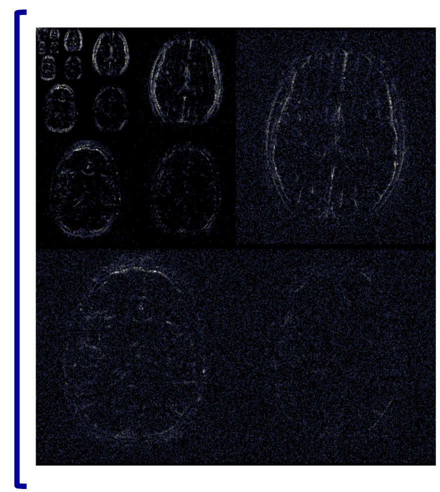
y



$=$



A



x

A few Fourier coefficients

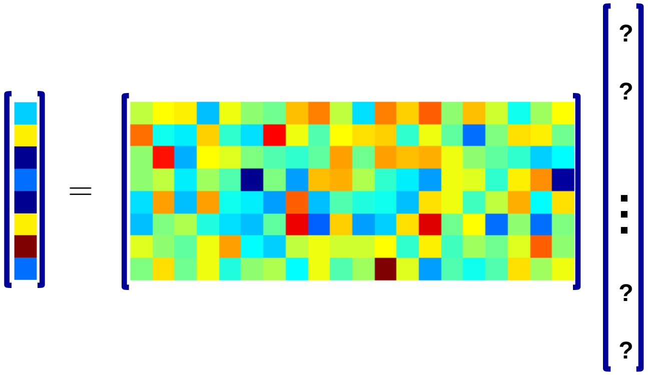
Wavelet coefficients

[Lustig, Donoho + Pauly ‘10] … brain image – Lustig ‘12

UNDERDETERMINED LINEAR SYSTEMS

Underdetermined system

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$



Compression

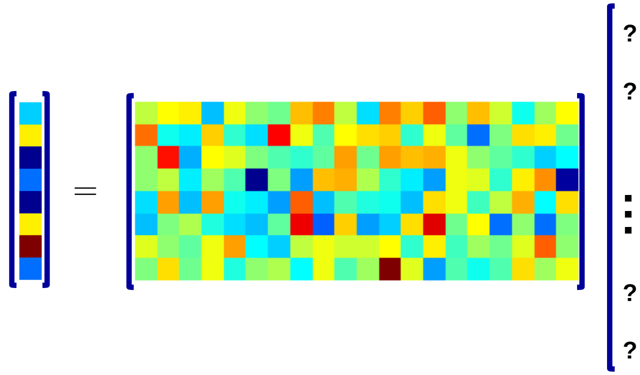
$\boldsymbol{y}$

Image to be compressed

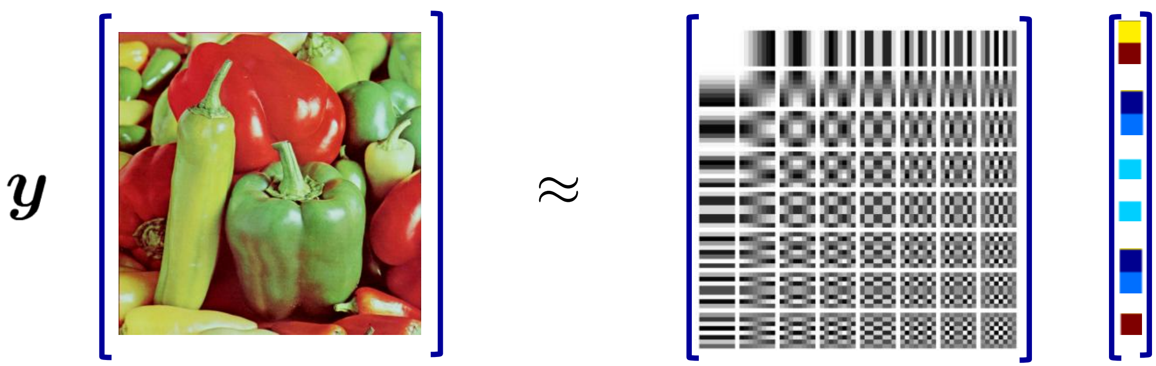
UNDERDETERMINED LINEAR SYSTEMS

Underdetermined system

$$y = Ax$$



Compression – JPEG



(Patches of) … input image

A DCT basis

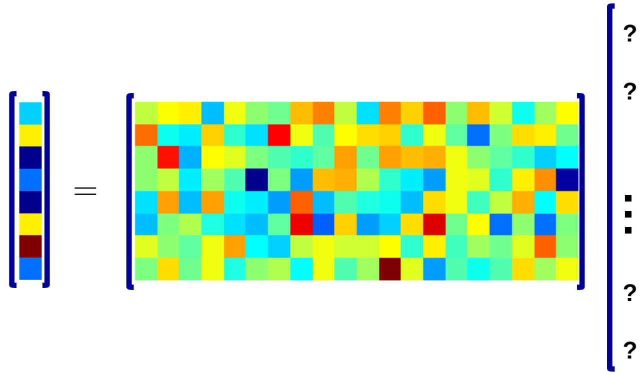
x coefficients

[Wallace '91]

UNDERDETERMINED LINEAR SYSTEMS

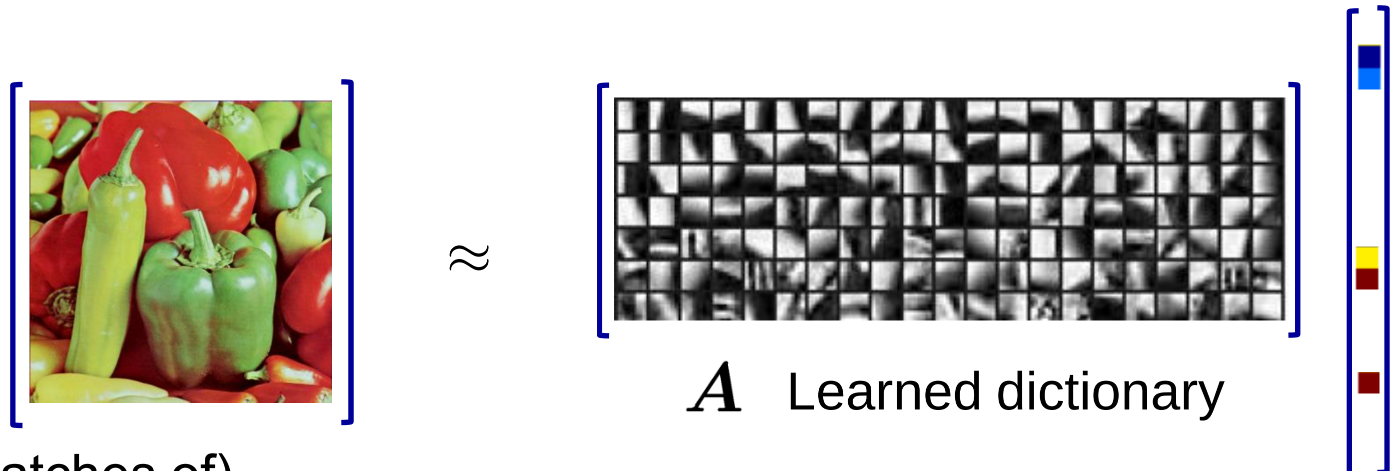
Underdetermined system

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$



Compression – Learned dictionary

$\boldsymbol{y}$



(Patches of) …
input image

$\boldsymbol{A}$ Learned dictionary

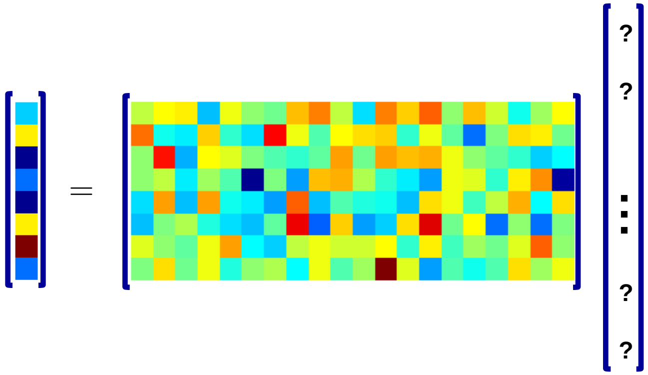
$\boldsymbol{x}$ coefficients

See [Elad+Bryt ’08], [Horev et. Al., ‘12] … Image: [Aharon+Elad ‘05]

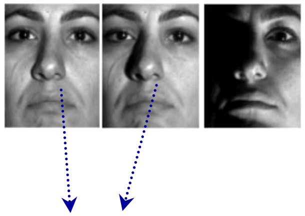
UNDERDETERMINED LINEAR SYSTEMS

Underdetermined system

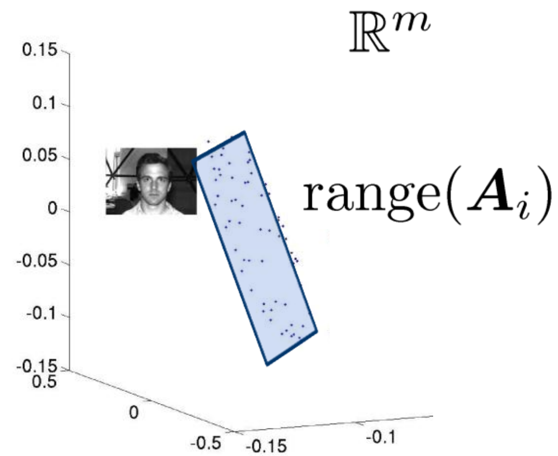
$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$



Recognition



$$\boldsymbol{A}_i = [\;|\;|\;\ldots\;] \in \mathbb{R}^{m \times n_i}$$



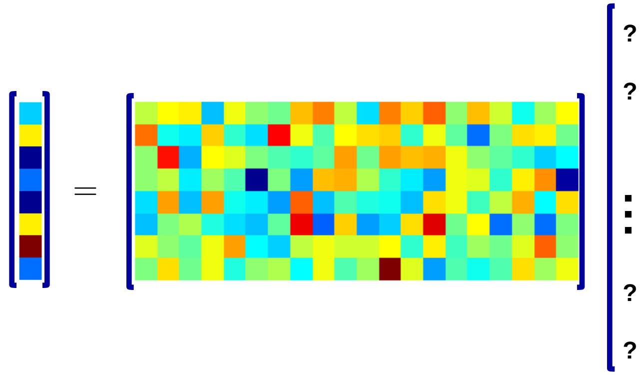
Linear subspace model for images of **same face** under **varying lighting.**

[Basri+Jacobs '03], [Ramamoorthi '03], [Belhumeur+Kriegman '96]

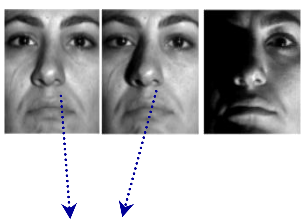
UNDERDETERMINED LINEAR SYSTEMS

Underdetermined system

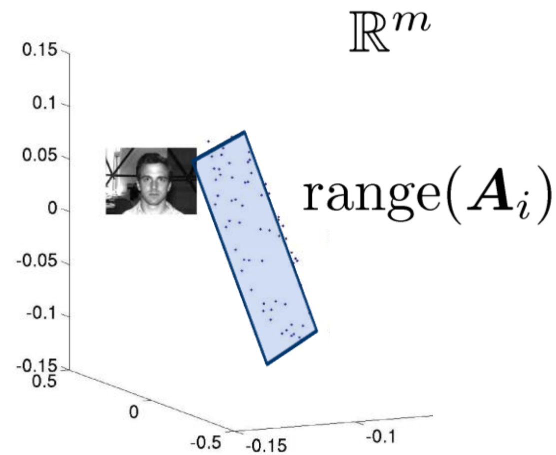
$$y = Ax$$



Recognition



$$A_i = [\ |\ |\ \ldots\ ] \in \mathbb{R}^{m \times n_i}$$

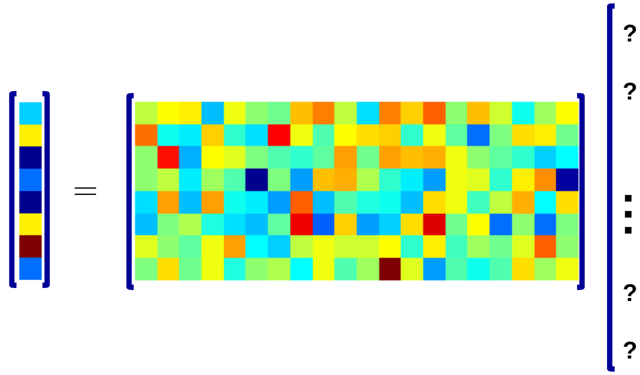


y  $\approx\ x_{i,1}$  $+\ x_{i,2}$  $+ \ldots +\ x_{i,n}$  $= A_i x_i$

UNDERDETERMINED LINEAR SYSTEMS

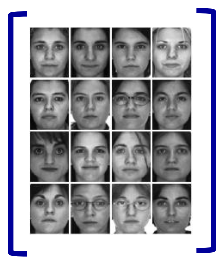
Underdetermined system

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

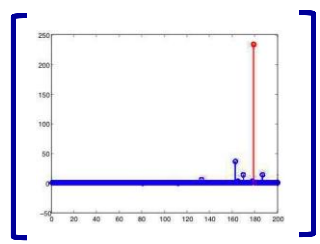


Recognition

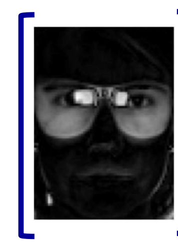
$=$



$\times$



$+$



$\boldsymbol{y} \in \mathbb{R}^m$

Test image

$\boldsymbol{A} = [\boldsymbol{A}_1 \mid \boldsymbol{A}_2 \mid \cdots \mid \boldsymbol{A}_k]$

Combined training dictionary

$\boldsymbol{x} \in \mathbb{R}^n$

coefficients

$\boldsymbol{e} \in \mathbb{R}^m$

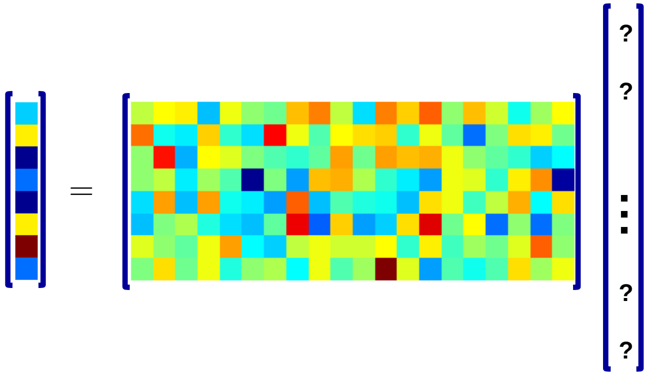
corruption, occlusion

[W., Yang, Ganesh, Sastry, Ma ’09]

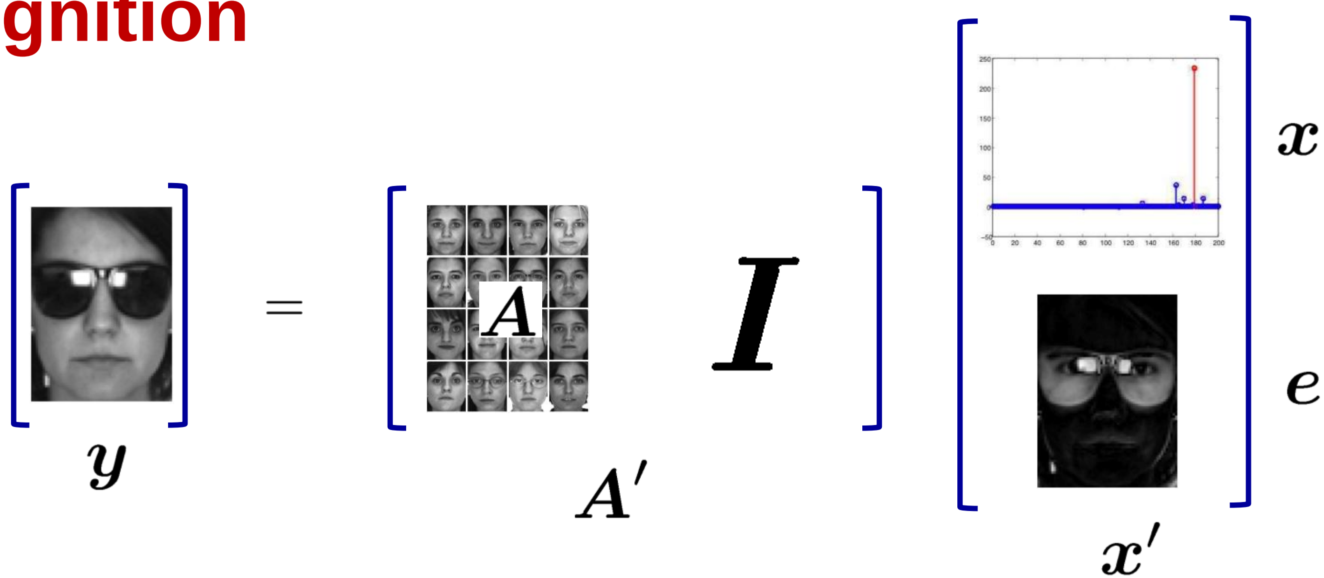
UNDERDETERMINED LINEAR SYSTEMS

Underdetermined system

$$y = Ax$$

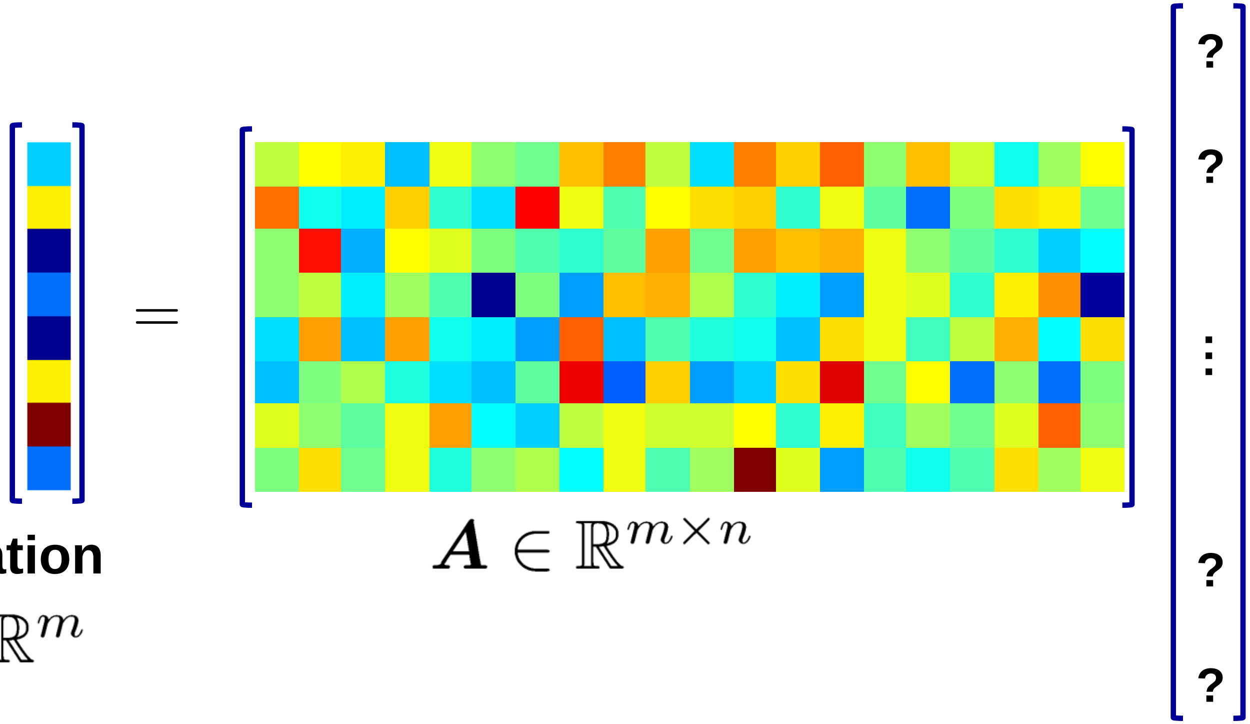


Recognition



One large underdetermined system: $y = A'x'$.

UNDERDETERMINED LINEAR SYSTEMS



Observation

$\boldsymbol{y} \in \mathbb{R}^m$

$\boldsymbol{A} \in \mathbb{R}^{m \times n}$

Unknown

$\boldsymbol{x} \in \mathbb{R}^n$

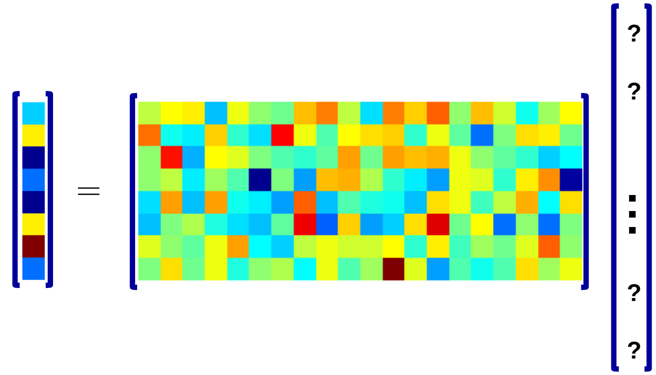
In all of these examples, $\underbrace{m}_{\#\texttt{observations}} \ll \underbrace{n}_{\#\texttt{unknowns}}$.

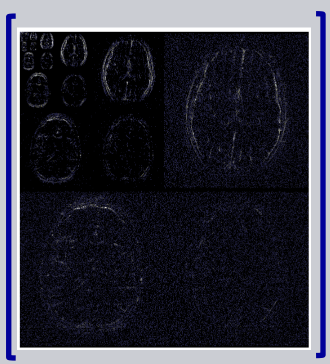
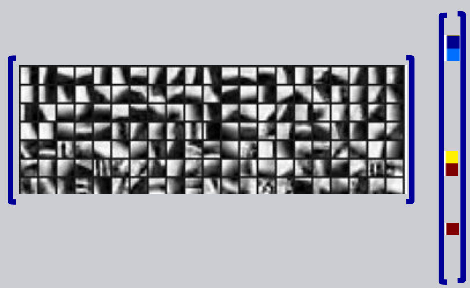
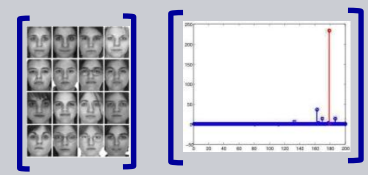
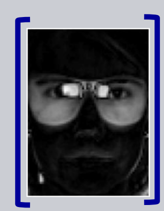
Solution is **not unique** … is there any hope?

WHAT DO WE KNOW ABOUT x?

Underdetermined system

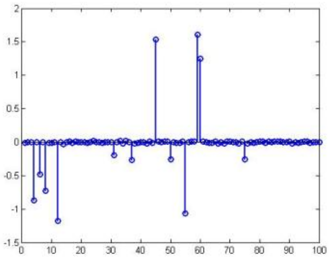
$$y = Ax$$



Signal acquisition	Image compression	Face Recognition
 $x^\star$ contains **just a few** significant wavelet coefficients.	 $x^\star$ uses **just a few** dictionary elements.	 $x^\star$ uses **just a few** training faces.  $e^\star$ corrects **a few** gross errors.

SPARSITY – More formally

A vector $\boldsymbol{x} \in \mathbb{R}^n$ is **sparse** if only a few entries are nonzero:

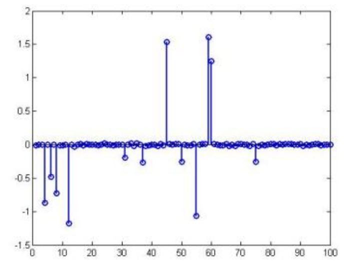


The **number of nonzeros** is called the ℓ^0**-"norm"** of $\boldsymbol{x}$:

$$\|\boldsymbol{x}\|_0 \doteq \#\{i \mid x_i \neq 0\}.$$

SPARSITY – More formally

A vector $\boldsymbol{x} \in \mathbb{R}^n$ is **sparse** if
only a few entries are nonzero:



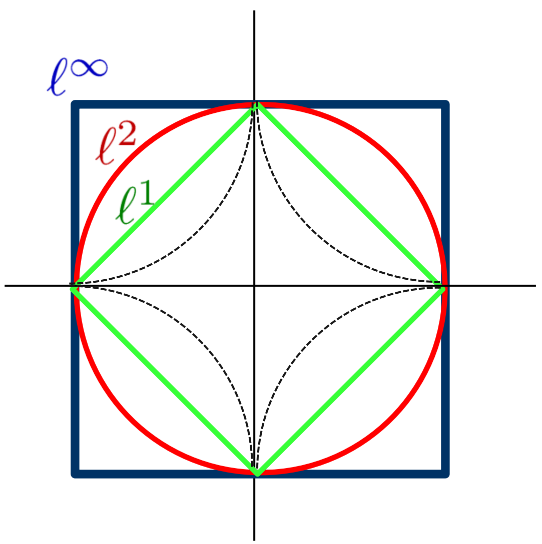
The **number of nonzeros** is called the ℓ^0**-"norm"** of $\boldsymbol{x}$:

$$\|\boldsymbol{x}\|_0 \doteq \#\{i \mid x_i \neq 0\}.$$

Geometrically

$$\|\boldsymbol{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$$

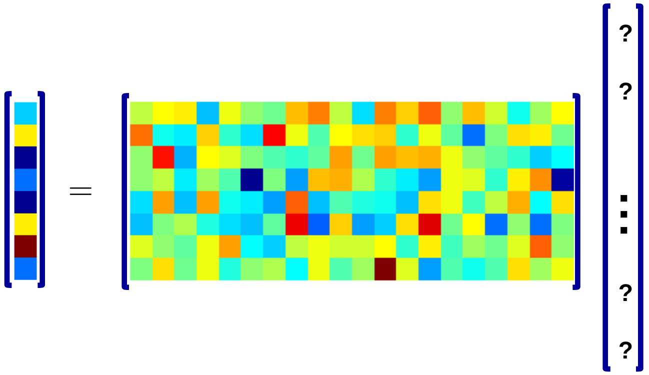
$$\|\boldsymbol{x}\|_0 = \lim_{p \searrow 0} \|\boldsymbol{x}\|_p^p.$$



THE SPARSEST SOLUTION

Underdetermined system

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$



Look for the sparsest $\boldsymbol{x}$ that agrees with our observation:

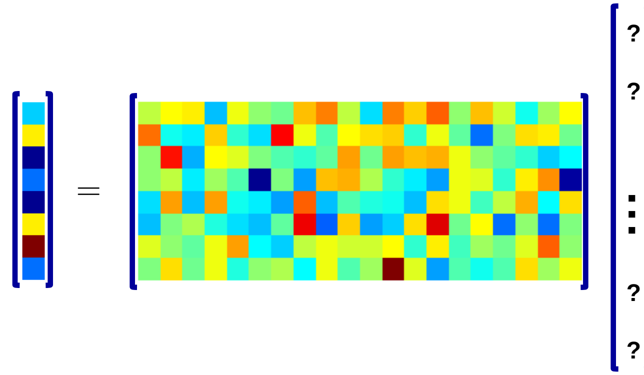
$$\text{minimize} \quad \|\boldsymbol{x}\|_0 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}.$$

[Demo]

THE SPARSEST SOLUTION

Underdetermined system

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$



Look for the sparsest $\boldsymbol{x}$ that agrees with our observation:

$$\text{minimize} \quad \|\boldsymbol{x}\|_0 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}.$$

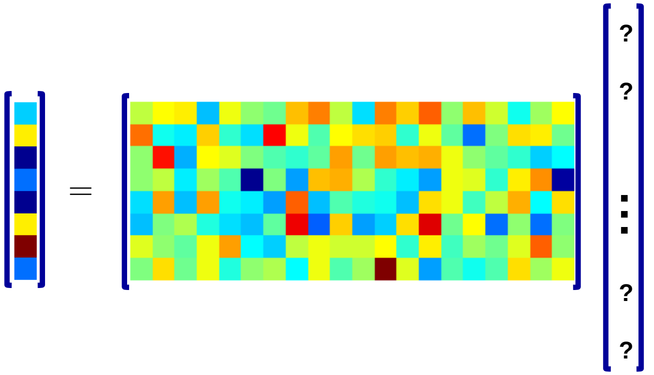
Theorem 1 (Gorodnitsky+Rao '97) . *Suppose* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0$, *and let* $k = \|\boldsymbol{x}_0\|_0$. *If* $\mathrm{null}(\boldsymbol{A})$ *contains no* $2k$*-sparse vectors,* $\boldsymbol{x}_0$ *is the unique optimal solution to*

$$\text{minimize } \|\boldsymbol{x}\|_0 \quad \textit{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}.$$

THE SPARSEST SOLUTION

Underdetermined system

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$



Look for the sparsest $\boldsymbol{x}$ that agrees with our observation:

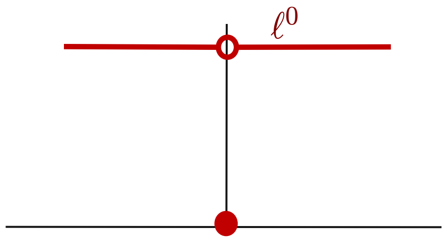
$$\xcancel{\text{minimize} \quad \|\boldsymbol{x}\|_0 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}.}$$

INTRACTABLE

RELAX!

$$\xcancel{\text{minimize} \quad \|\boldsymbol{x}\|_0 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}.}$$

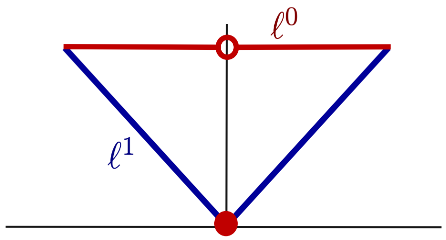
The cardinality $\|\boldsymbol{x}\|_0$ is **nonconvex**:



RELAX!

$$\cancel{\text{minimize} \quad \|\boldsymbol{x}\|_0 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}.}$$

The cardinality $\|\boldsymbol{x}\|_0$ is **nonconvex**:



Its **convex envelope*** is

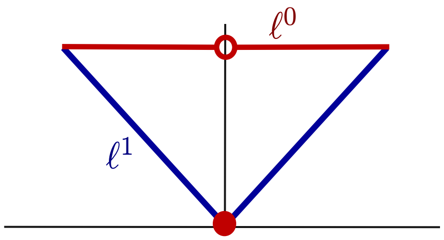
the ℓ^1 **norm:** $\|\boldsymbol{x}\|_1 = \sum_i |x_i|$

* Over the set $\{\boldsymbol{x} \mid |x_i| \leq 1 \; \forall \; i\}$.

RELAX!

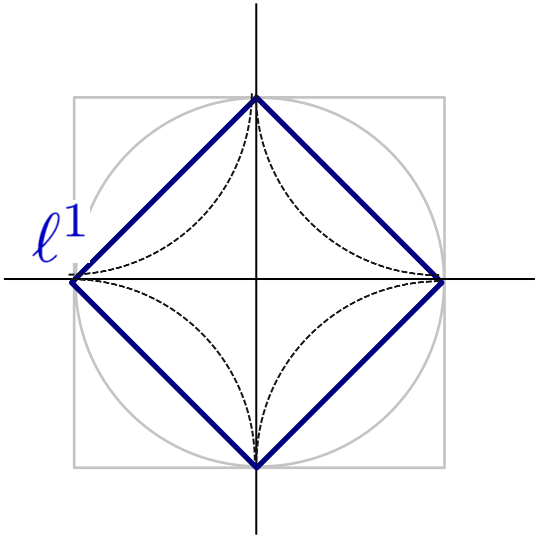
$$\xcancel{\text{minimize} \quad \|\boldsymbol{x}\|_0 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}.}$$

The cardinality $\|\boldsymbol{x}\|_0$ is **nonconvex**:



Its **convex envelope*** is

the ℓ^1 **norm:** $\|\boldsymbol{x}\|_1 = \sum_i |x_i|$



* Over the set $\{\boldsymbol{x} \mid |x_i| \le 1 \ \forall\ i\}$.

RELAX!

$$\text{minimize} \quad \|\boldsymbol{x}\|_0 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}.$$

NP-hard, hard to appx.
[Natarjan ’95],
[Amaldi+Kann ‘97]

⬇ ⬇

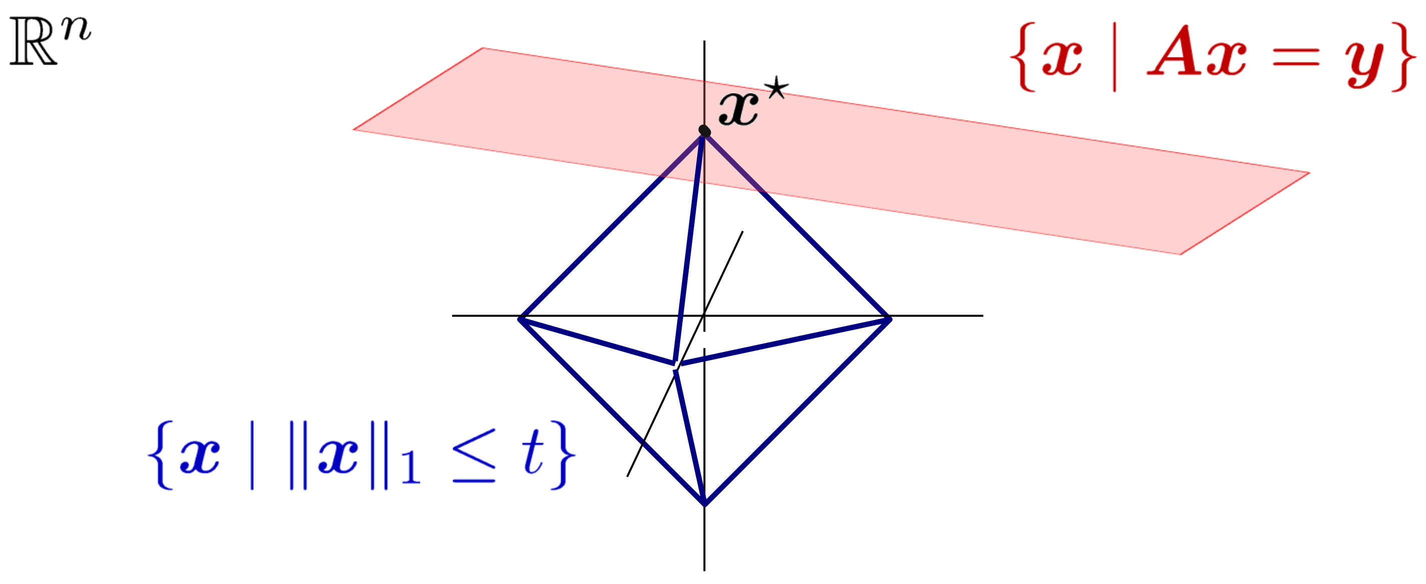
$$\text{minimize} \quad \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}.$$

Efficiently solvable
– Lecture 2!

Have we lost anything? [demo]

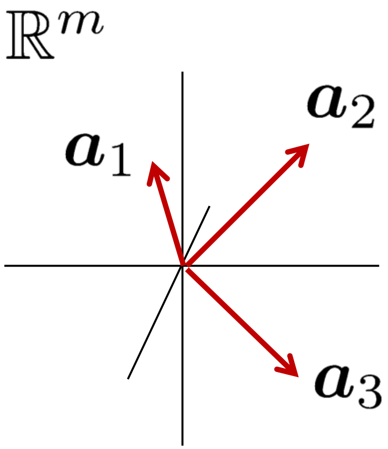
WHY DOES THIS WORK? Geometric intuition

$$\text{minimize} \quad \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}.$$



WHY DOES THIS WORK? More formally…

We see: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \sum_{i \in \mathrm{supp}(\boldsymbol{x})} \boldsymbol{a}_i x_i$.



Intuition: Recovering $\boldsymbol{x}$ is “easier” if the $\boldsymbol{a}_i$ are not too similar…

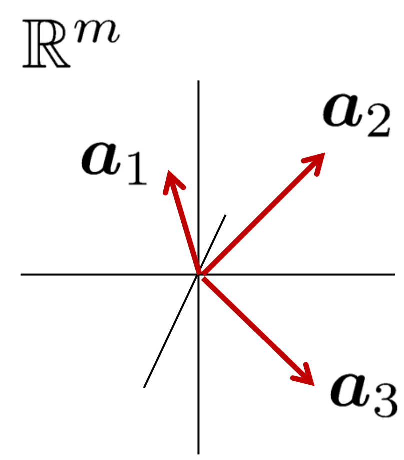
Mutual coherence $\mu(\boldsymbol{A}) \doteq \max_{i \neq j} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle|$.

Smaller is better!

WHY DOES THIS WORK? More formally…

Mutual coherence

$$\mu(\boldsymbol{A}) \doteq \max_{i \neq j} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle|.$$



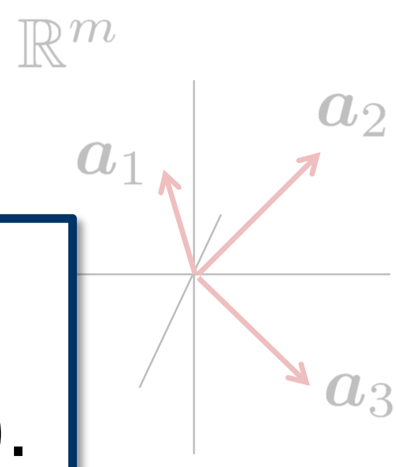
Theorem 2 (Gribonval+Nielsen '03, Donoho+Elad '03) *Suppose* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0$ *with*

$$\|\boldsymbol{x}_0\|_0 < \tfrac{1}{2}(1 + 1/\mu(\boldsymbol{A})).$$

Then $\boldsymbol{x}_0$ *is the unique optimal solution to*

$$\text{minimize } \|\boldsymbol{x}\|_1 \quad \textit{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}.$$

WHY DOES THIS WORK? More formally…



Mutual coherence

The target solution $\boldsymbol{x}_0$ is **sufficiently structured** (sparse!).

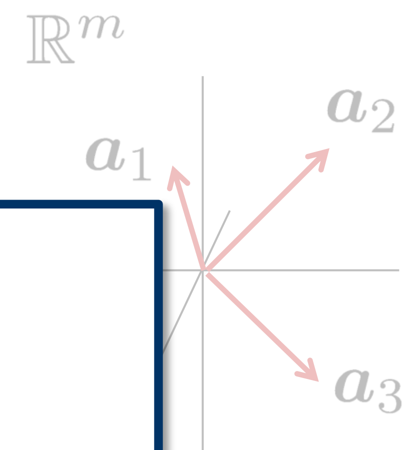
Theorem 2 (Gribonval+Nielsen '03, Donoho+Elad '03) *Suppose* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0$ *with*

$$\|\boldsymbol{x}_0\|_0 < \tfrac{1}{2}(1 + 1/\mu(\boldsymbol{A})).$$

Then $\boldsymbol{x}_0$ *is the unique optimal solution to*

$$\text{minimize } \|\boldsymbol{x}\|_1 \quad \textit{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}.$$

WHY DOES THIS WORK? More formally…



Mutual coherence

The matrix $\boldsymbol{A}$ is **incoherent** – and so, preserves sparse $\boldsymbol{x}$.

Theorem 2 (Gribonval+Nielsen '03, Donoho+Elad '03) *Suppose* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0$ *with*

$$\|\boldsymbol{x}_0\|_0 < \tfrac{1}{2}(1 \boxed{+ 1/\mu(\boldsymbol{A}))}.$$

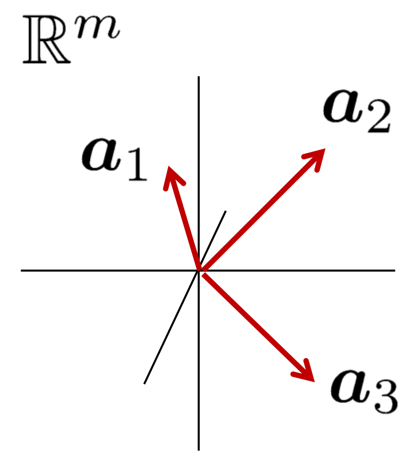
Then $\boldsymbol{x}_0$ *is the unique optimal solution to*

$$\text{minimize } \|\boldsymbol{x}\|_1 \quad \textit{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}.$$

WHY DOES THIS WORK? More formally…

Mutual coherence

$$\mu(\boldsymbol{A}) \doteq \max_{i \neq j} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle|.$$



Theorem 2 (Gribonval+Nielsen '03, Donoho+Elad '03) *Suppose* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0$ *with*

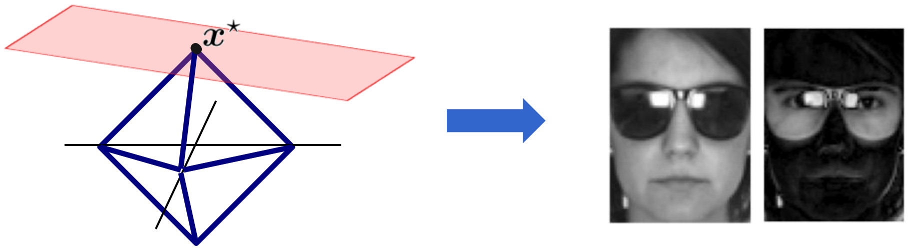
$$\|\boldsymbol{x}_0\|_0 < \tfrac{1}{2}(1 + 1/\mu(\boldsymbol{A})).$$

Then $\boldsymbol{x}_0$ *is the unique optimal solution to*

$$\text{minimize } \|\boldsymbol{x}\|_1 \quad \textit{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}.$$

WHY CARE ABOUT THE THEORY?

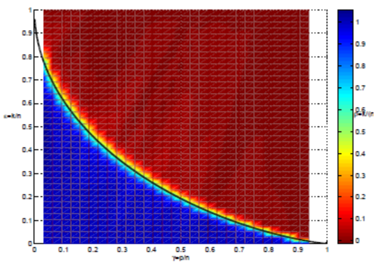
Motivates applications



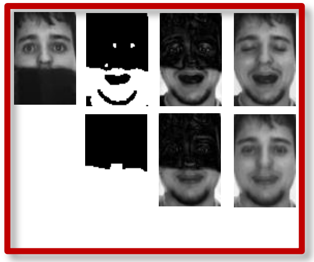
… but be careful: sometimes need to modify basic formulation [Lecture 3].

Template for stronger results

… predictions can be very sharp in high dimensions.

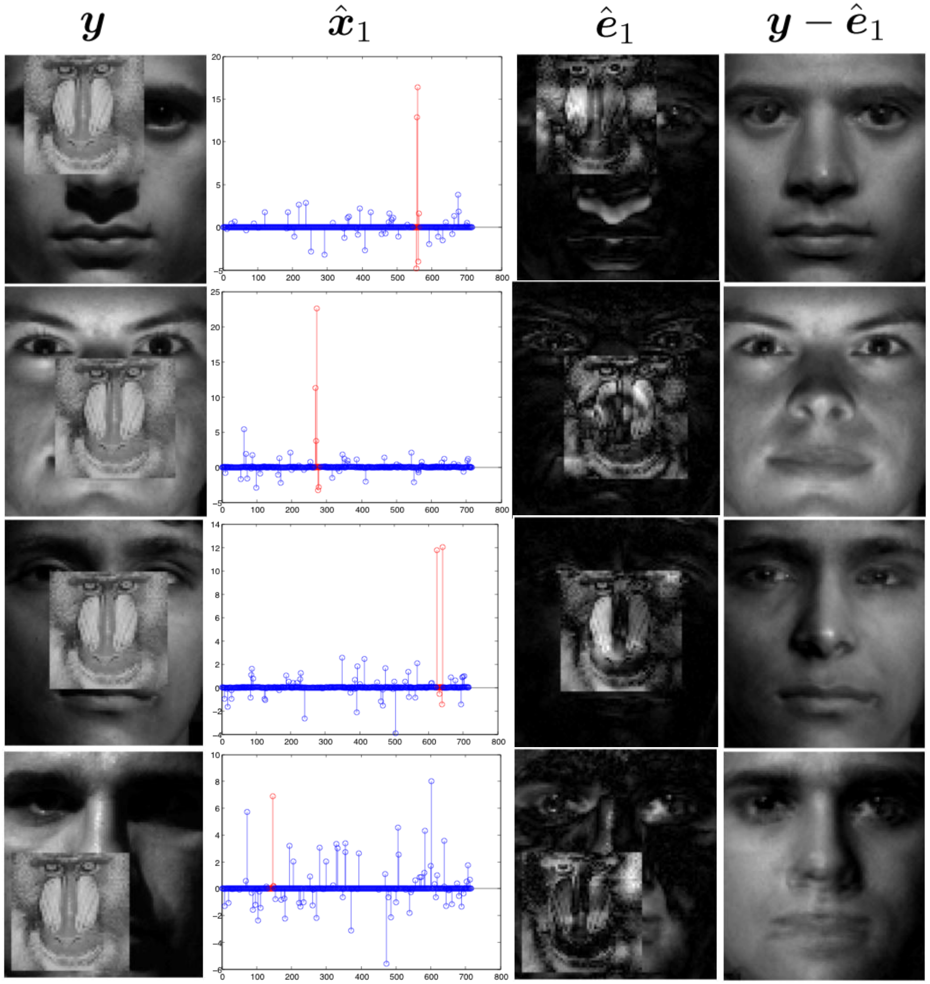


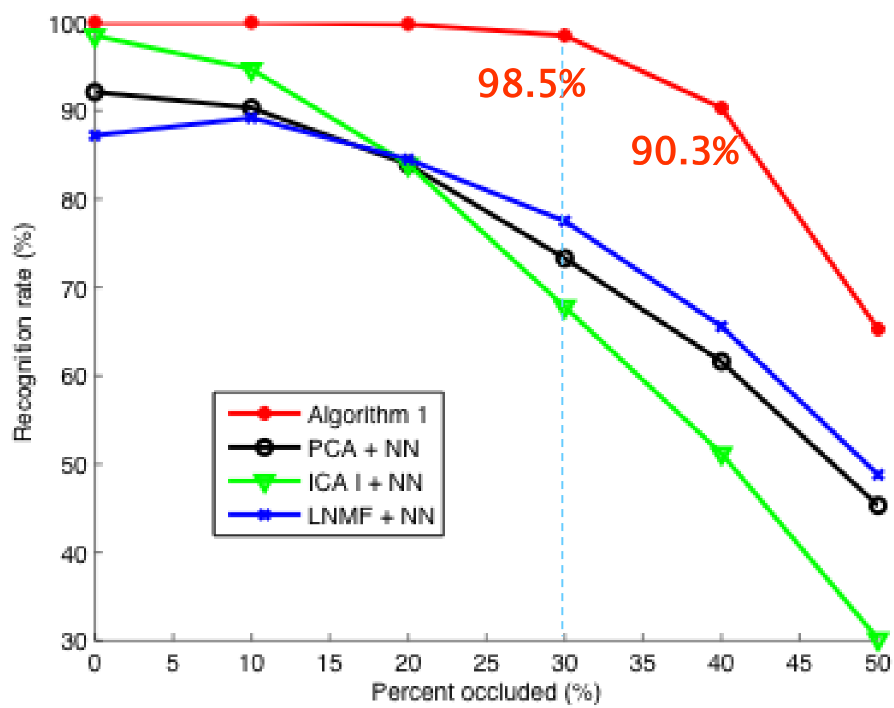
Generalizes to many other types
of low-dimensional structure



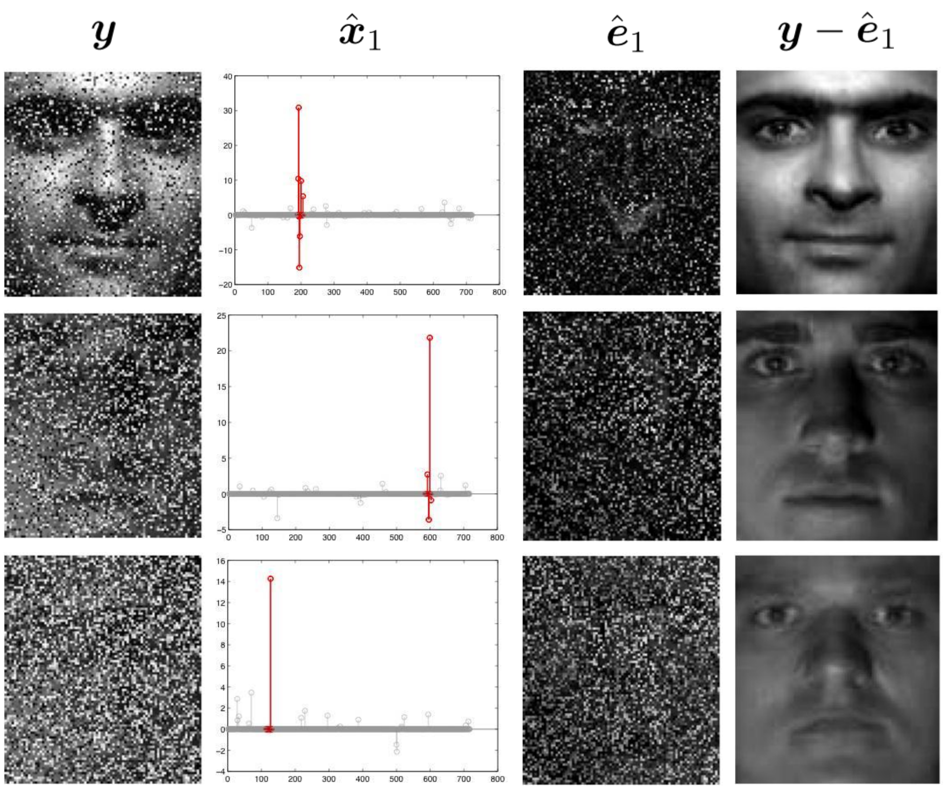
… structured sparsity [Lecture 2], low-rank recovery [later, Lecture 3].

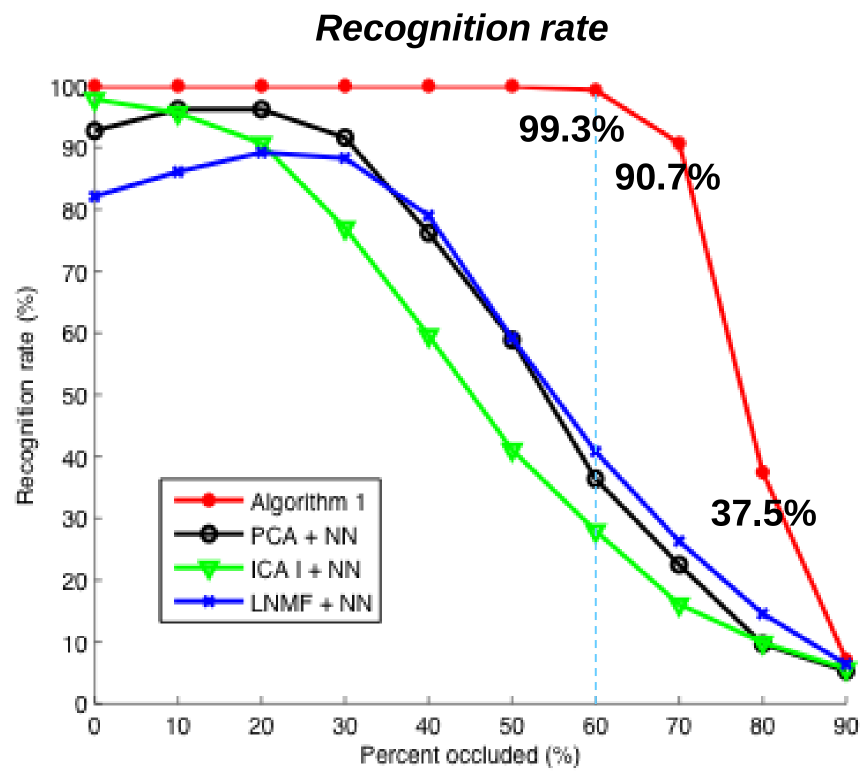
THEORY TO APPLICATION – Face Recognition





THEORY TO APPLICATION – Face Recognition

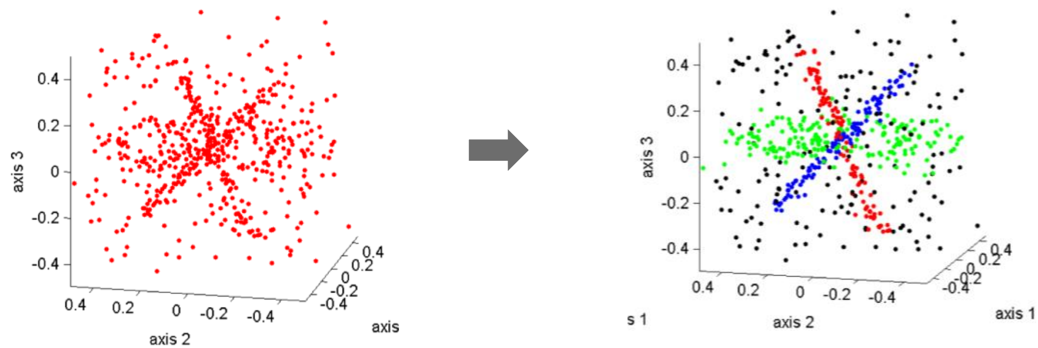




More practicalities in Lecture 3…

THEORY TO APPLICATION – Subspace Segmentation

$$\text{Given } \boldsymbol{Y} = [\boldsymbol{y}_1 \mid \cdots \mid \boldsymbol{y}_N] \subset \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_k, \text{ determine } (\mathcal{S}_i).$$



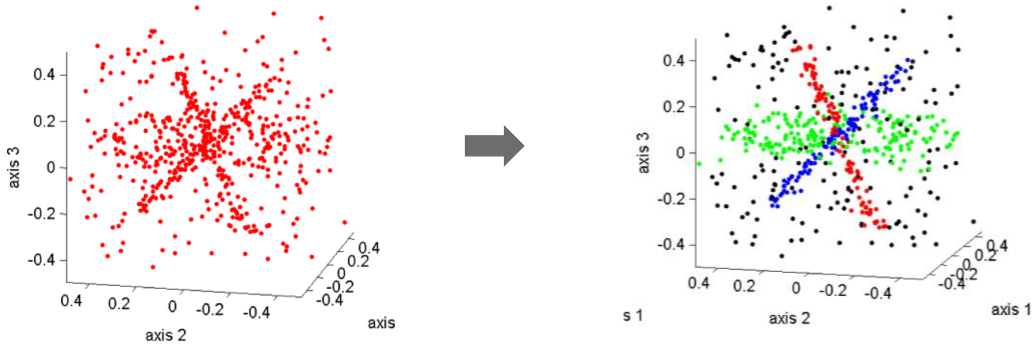
Data $\boldsymbol{Y}$

Segmentation

Applications include
image segmentation,
motion segmentation,
hybrid system identification,
and more.

THEORY TO APPLICATION – Subspace Segmentation

Given $\boldsymbol{Y} = [\boldsymbol{y}_1 \mid \cdots \mid \boldsymbol{y}_N] \subset \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_k$, determine $(\mathcal{S}_i)$.



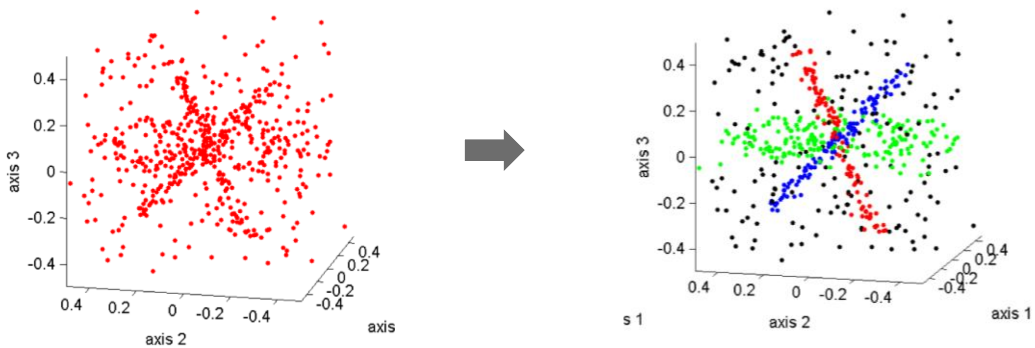
Data $\boldsymbol{Y}$ Segmentation

Each $\boldsymbol{y}_i \in \mathcal{S}_j$ can be expressed as a linear combination of just $\dim(\mathcal{S}_j)$ other points $\boldsymbol{y}'_i \in \mathcal{S}_j$.

[Elhamifar+Vidal, ’09]

THEORY TO APPLICATION – Subspace Segmentation

Given $\boldsymbol{Y} = [\boldsymbol{y}_1 \mid \cdots \mid \boldsymbol{y}_N] \subset \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_k$, determine $(\mathcal{S}_i)$.



Data $\boldsymbol{Y}$ Segmentation

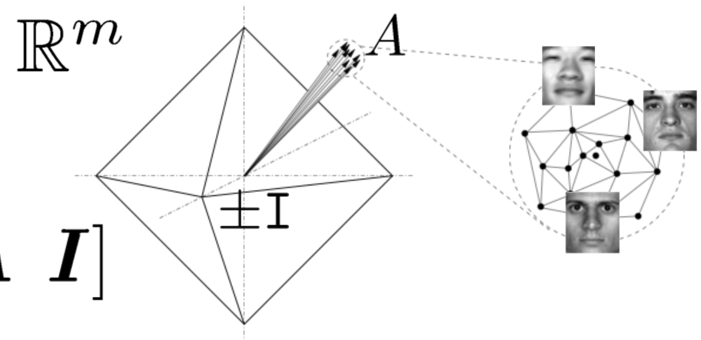
Each $\boldsymbol{y}_i \in \mathcal{S}_j$ can be expressed as a linear combination of just $\dim(\mathcal{S}_j)$ other points $\boldsymbol{y}_i' \in \mathcal{S}_j$.

$$\text{minimize} \quad \|\boldsymbol{X}\|_1 \quad \text{subject to} \quad \boldsymbol{Y}\boldsymbol{X} = \boldsymbol{Y}, \ \operatorname{diag}(\boldsymbol{X}) = 0.$$

[Elhamifar+Vidal, '09]

THEORY AND PRACTICE – Faces and Subspaces

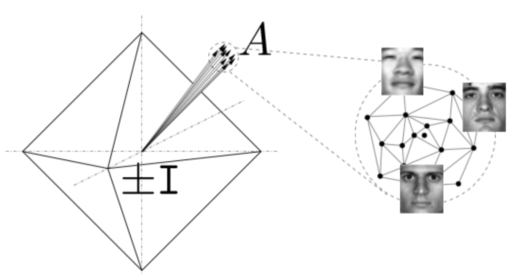
In both applications, $\boldsymbol{A}$ can be coherent…



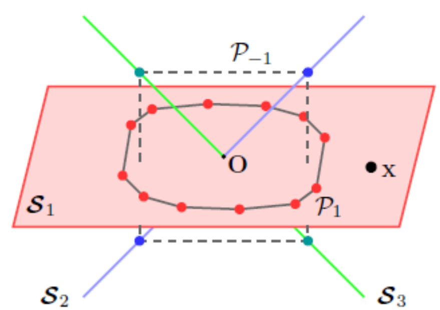
$$\boldsymbol{A}' = [\boldsymbol{A} \;\; \boldsymbol{I}]$$

… ℓ^1 still exhibits **exact recovery** when $\boldsymbol{x}_0$ is structured.

Theory extends by considering **problem-specific geometry**



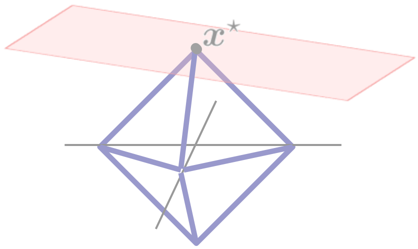
[W.+Ma ’10]



[Elhamifar+Vidal, ‘12]
[Soltanolkotabi and Candes, ‘11].

WHY CARE ABOUT THE THEORY?

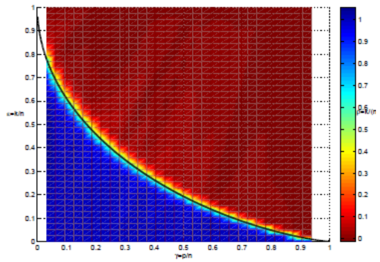
Motivates applications



… but be careful: sometimes need to modify basic formulation [Lecture 3].

Template for stronger results

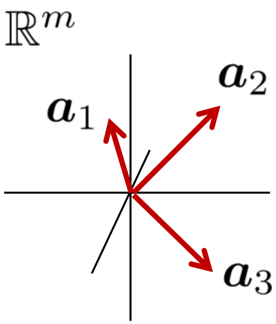
… predictions can be very sharp in high dimensions.



Generalizes to many other types
of low-dimensional structure

… structured sparsity [Lecture 2], low-rank recovery [later, Lecture 3].

LIMITATIONS OF COHERENCE?

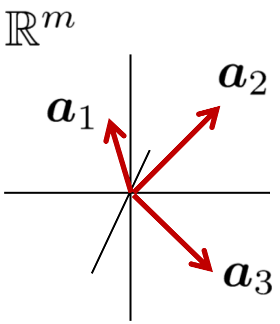


For any $m \times n$ $\boldsymbol{A}$, $\mu(\boldsymbol{A}) \geq \sqrt{\frac{n-m}{m(n-1)}}$.

Prev. result therefore requires

$$\|\boldsymbol{x}_0\|_0 \; < \; \tfrac{1}{2}(1 + \mu(\boldsymbol{A})^{-1}) \; = \; O(\sqrt{m})$$

LIMITATIONS OF COHERENCE?

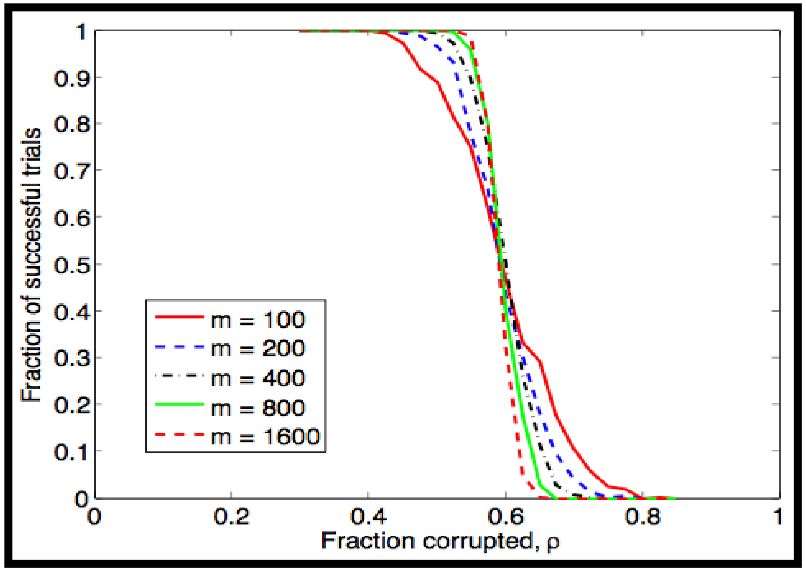


For any $m \times n$ $\boldsymbol{A}$, $\mu(\boldsymbol{A}) \geq \sqrt{\frac{n-m}{m(n-1)}}$.

Prev. result therefore requires

$$\|\boldsymbol{x}_0\|_0 \ < \ \tfrac{1}{2}(1 + \mu(\boldsymbol{A})^{-1}) \ = \ O(\sqrt{m})$$

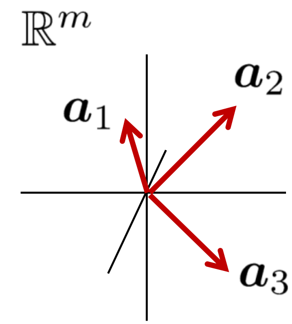
Truth is often **much better**:



Plot: Fraction of correct recovery
vs. fraction of nonzeros $\|\boldsymbol{x}_0\|_0/m$

LIMITATIONS OF COHERENCE?

For any $m \times n$ $\boldsymbol{A}$, $\mu(\boldsymbol{A}) \geq \sqrt{\frac{n-m}{m(n-1)}}$.



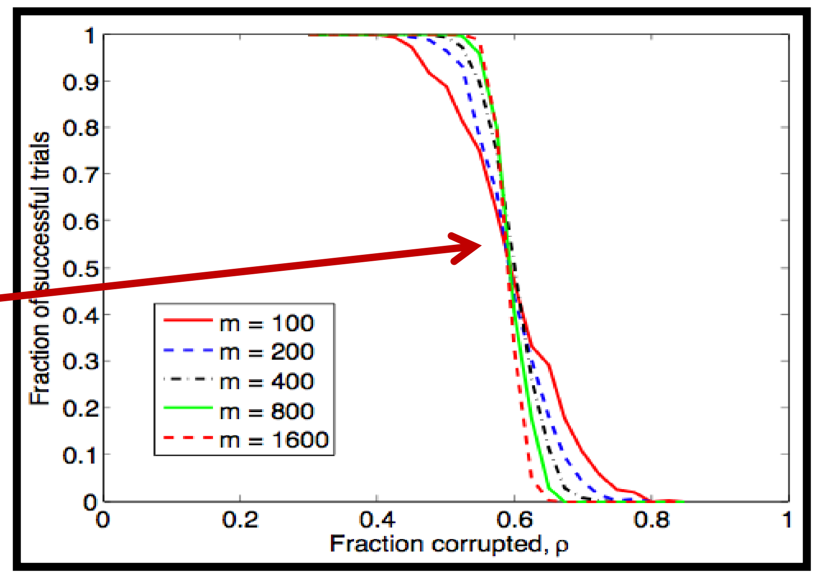
Prev. result therefore requires

$$\|\boldsymbol{x}_0\|_0 \; < \; \tfrac{1}{2}\left(1 + \mu(\boldsymbol{A})^{-1}\right) \; = \; O(\sqrt{m})$$

Truth is often **much better**:

Phase transition at

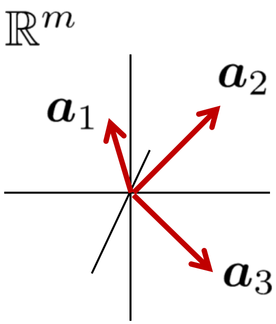
$$\|\boldsymbol{x}_0\|_0 = \alpha^\star m$$



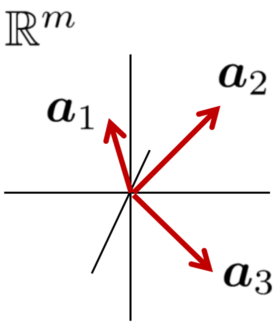
Plot: Fraction of correct recovery
vs. fraction of nonzeros $\|\boldsymbol{x}_0\|_0 / m$

STRENGTHENING THE BOUND – the RIP

Incoherence: **Each pair** $\boldsymbol{A}_{i,j} = [\boldsymbol{a}_i \mid \boldsymbol{a}_j]$ spread.



STRENGTHENING THE BOUND – the RIP



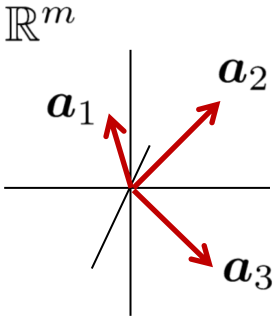
Incoherence: **Each pair** $\boldsymbol{A}_{i,j} = [\boldsymbol{a}_i \mid \boldsymbol{a}_j]$ spread.

Generalize to **subsets of size** k **:**

$\boldsymbol{A}_I$ well-spread (almost orthonormal) for all I of size k

$\implies$ all k-sparse $\boldsymbol{x}$, $\|\boldsymbol{A}\boldsymbol{x}\|_2 \approx \|\boldsymbol{x}\|_2$

STRENGTHENING THE BOUND – the RIP



Incoherence: **Each pair** $\boldsymbol{A}_{i,j} = [\boldsymbol{a}_i \mid \boldsymbol{a}_j]$ spread.

Generalize to **subsets of size** k **:**

$\boldsymbol{A}_I$ well-spread (almost orthonormal) for all I of size k

$\implies$ all k-sparse $\boldsymbol{x}$, $\|\boldsymbol{A}\boldsymbol{x}\|_2 \approx \|\boldsymbol{x}\|_2$

$\boldsymbol{A}$ satisfies the **Restricted Isometry Property** of order k, with constant δ if for all k-sparse $\boldsymbol{x}$,

$$(1-\delta)\|\boldsymbol{x}\|_2^2 \;\le\; \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \;\le\; (1+\delta)\|\boldsymbol{x}\|_2^2.$$

IMPLICATIONS OF RIP

Good sparse recovery

Theorem 2 (Candès+Tao '05, Candès '08)
Suppose $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0$ *with*

$$\delta_{2\|\boldsymbol{x}_0\|_0} < \sqrt{2} - 1.$$

Then $\boldsymbol{x}_0$ *is the unique optimal solution to*

$$\text{minimize } \|\boldsymbol{x}\|_1 \quad \textit{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}.$$

IMPLICATIONS OF RIP

Good sparse recovery

Theorem 2 (Candès+Tao '05, Candès '08) *Suppose* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0$ *with*

$$\delta_{2\|\boldsymbol{x}_0\|_0} < \sqrt{2} - 1.$$

Then $\boldsymbol{x}_0$ *is the unique optimal solution to*

$$\text{minimize } \|\boldsymbol{x}\|_1 \quad \textit{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}.$$

Again, if … x_0 **is "structured" and** A **is "nice"**

we exactly recover x_0 **.**

Compare condition to condition $\|x_0\|_0 < \frac{1}{2}(1 + \mu(\boldsymbol{A})^{-1})$**.**

IMPLICATIONS OF RIP

Random A are great:

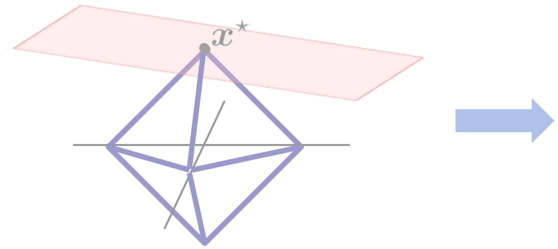
If $\boldsymbol{A} \sim_{iid} \mathcal{N}(0, m^{-1/2})$, then $\boldsymbol{A}$ has RIP of order k with high probability, when $m \geq Ck \log(n/k)$.

For random $\boldsymbol{A}$. ℓ^1 works even when $\|\boldsymbol{x}_0\|_0 \sim m$.

Useful property for designing sampling operators (***Compressed sensing***).

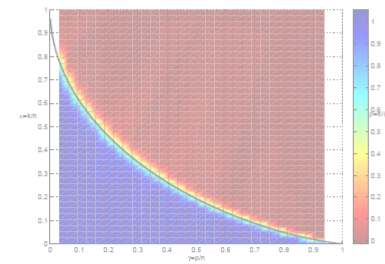
WHY CARE ABOUT THE THEORY?

Motivates applications



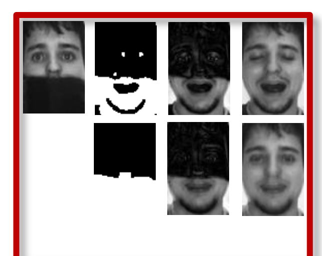
… but be careful: sometimes need to modify basic formulation [Lecture 3].

Template for stronger results



… predictions can be very sharp in high dimensions.

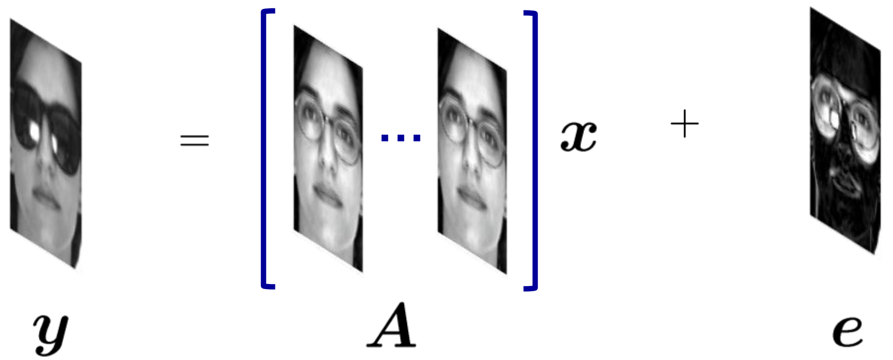
Generalizes to many other types of low-dimensional structure



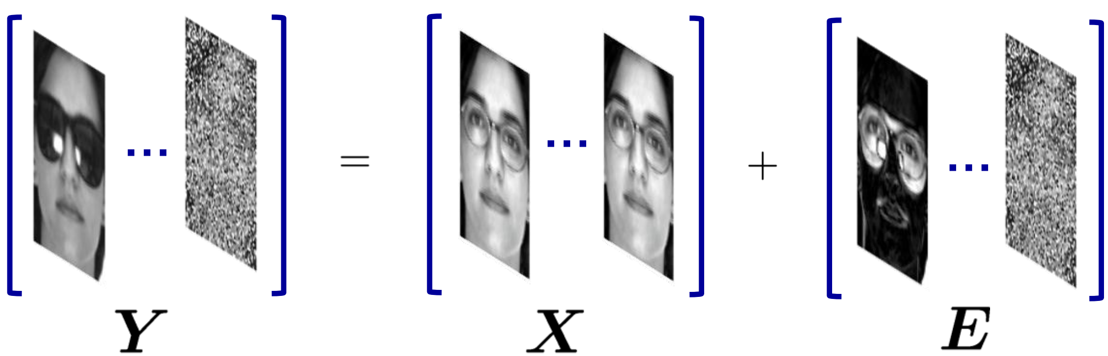
… structured sparsity [Lecture 2], low-rank recovery [later, Lecture 3].

GENERALIZATIONS – From Sparse to Low-Rank

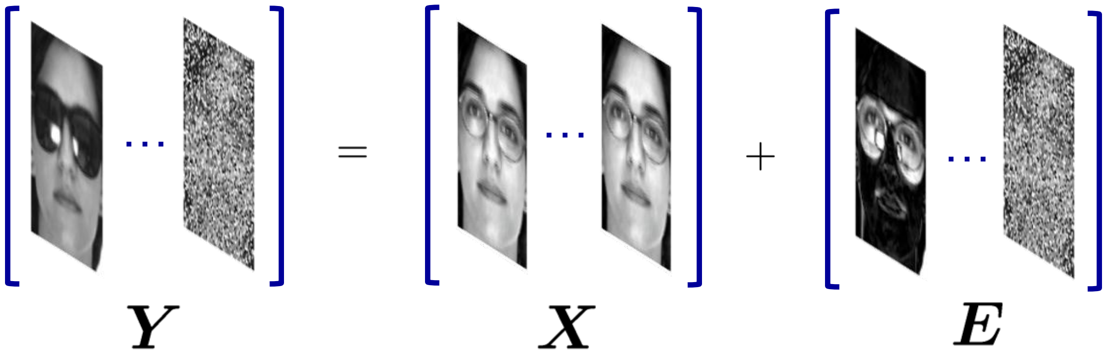
So far: Recovering a single sparse vector:



Next: Recovering low-rank matrix (many correlated vectors):



FORMULATION – Robust PCA?



Given $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{E}$, with $\boldsymbol{X}$ low-rank, $\boldsymbol{E}$ sparse, recover $\boldsymbol{X}$.

Numerous approaches to **robust PCA** in the literature:

- Multivariate trimming *[Gnanadeskian + Kettering '72]*
- Random sampling *[Fischler + Bolles '81]*
- Alternating minimization *[Ke + Kanade '03]*
- Influence functions *[de la Torre + Black '03]*

Can we give an efficient, provably correct algorithm?

RELATED SOLUTIONS – Matrix recovery

Classical PCA/SVD – low rank + noise *[Hotelling '35, Karhunen+Loeve '72,…]*

Given $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z},$ recover $\boldsymbol{X}$.

Stable, efficient algorithm, theoretically optimal $\rightarrow$ huge impact

Matrix Completion – low rank, missing data

[Candès + Recht '08,
Candès + Tao '09,
Keshevan, Oh, Montanari '09,
Gross '09,
Ravikumar and Wainwright '10]

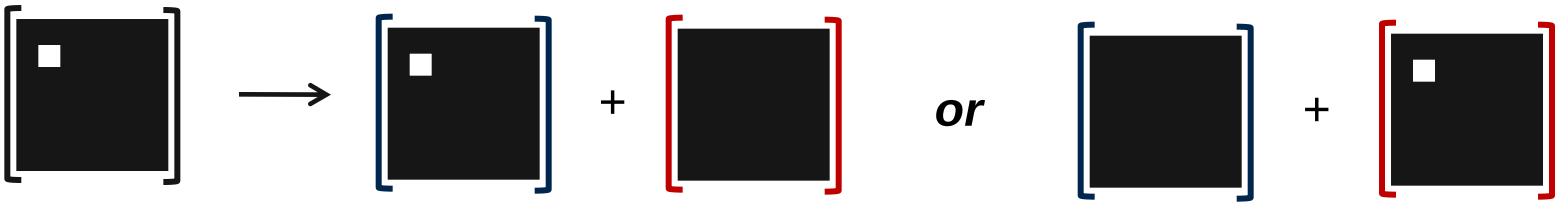
From $\boldsymbol{Y} = \mathcal{P}_\Omega[\boldsymbol{X}],$ recover $\boldsymbol{X}$.

Increasingly well-understood; solvable if $\boldsymbol{X}$ is low rank and Ω large enough.

Our problem, with $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{E}$, looks more difficult…

WHY IS THE PROBLEM HARD?

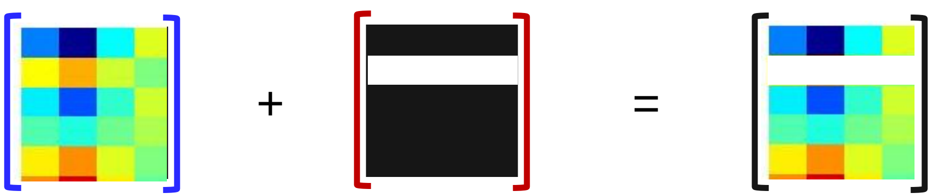
Some very sparse matrices are also low-rank:



$\boldsymbol{Y} = \mathbf{1}_{ij}$ $\boldsymbol{X} = \mathbf{1}_{ij}$ $\boldsymbol{E} = \mathbf{0}$ $\boldsymbol{X} = \mathbf{0}$ $\boldsymbol{E} = \mathbf{1}_{ij}$

*Can we recover $\boldsymbol{X}$ that are **incoherent** with the standard basis?*

Certain sparse error patterns $\boldsymbol{E}$ make recovering $\boldsymbol{X}$ impossible:



$\boldsymbol{X}$ $\boldsymbol{E} = \boldsymbol{e}_i \boldsymbol{v}^*$ $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{E}$

*Can we correct $\boldsymbol{E}$ whose support is not **adversarial**?*

WHEN IS THERE HOPE? Again, (in)coherence

Can we recover $\boldsymbol{X}$ that are ***incoherent*** *with the standard basis from* ***almost all*** *errors $\boldsymbol{E}$?*

Incoherence condition on singular vectors, **singular values arbitrary:**

Singular vectors of $\boldsymbol{X}$ not too spiky: $\begin{cases} \max_i \|\boldsymbol{U}_i\|^2 \le \mu r/m. \\ \max_i \|\boldsymbol{V}_i\|^2 \le \mu r/n. \end{cases}$

not too cross-correlated: $\|\boldsymbol{U}\boldsymbol{V}^*\|_\infty \le \sqrt{\mu r/mn}$

Uniform model on error support, **signs and magnitudes arbitrary:**

$$\mathrm{support}(\boldsymbol{E}) \sim \mathrm{uni}\binom{[m]\times[n]}{\rho mn}$$

Incoherence condition: [Candès + Recht ‘08]

… AND HOW SHOULD WE SOLVE IT?

Naïve optimization approach

Look for a low-rank $\boldsymbol{X}$ that agrees with the data up to some sparse error $\boldsymbol{E}$:

$$\min \ \operatorname{rank}(\boldsymbol{X}) + \gamma \|\boldsymbol{E}\|_0 \quad \text{subj} \quad \boldsymbol{X} + \boldsymbol{E} = \boldsymbol{Y}.$$

$$\operatorname{rank}(\boldsymbol{X}) = \#\{\sigma_i(\boldsymbol{X}) \neq 0\}. \qquad \|\boldsymbol{E}\|_0 = \#\{\boldsymbol{E}_{ij} \neq 0\}.$$

… AND HOW SHOULD WE SOLVE IT?

Naïve optimization approach

Look for a low-rank $\boldsymbol{X}$ that agrees with the data up to some sparse error $\boldsymbol{E}$:

$$\min \ \operatorname{rank}(\boldsymbol{X}) + \gamma\|\boldsymbol{E}\|_0 \quad \text{subj} \quad \boldsymbol{X} + \boldsymbol{E} = \boldsymbol{Y}.$$

$$\operatorname{rank}(\boldsymbol{X}) = \#\{\sigma_i(\boldsymbol{X}) \neq 0\}. \qquad \|\boldsymbol{E}\|_0 = \#\{\boldsymbol{E}_{ij} \neq 0\}.$$

INTRACTABLE

… AND HOW SHOULD WE SOLVE IT?

Naïve optimization approach

Look for a low-rank $\boldsymbol{X}$ that agrees with the data up to some sparse error $\boldsymbol{E}$:

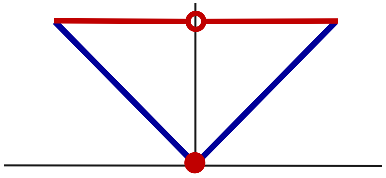
$$\min\ \operatorname{rank}(\boldsymbol{X}) + \gamma\|\boldsymbol{E}\|_0 \quad \text{subj} \quad \boldsymbol{X} + \boldsymbol{E} = \boldsymbol{Y}.$$

Convex relaxation

$$\operatorname{rank}(\boldsymbol{X}) = \#\{\sigma_i(\boldsymbol{X}) \neq 0\}. \qquad \|\boldsymbol{E}\|_0 = \#\{\boldsymbol{E}_{ij} \neq 0\}.$$

$$\downdownarrows \qquad\qquad\qquad \downdownarrows$$

$$\|\boldsymbol{X}\|_* = \textstyle\sum_i \sigma_i(\boldsymbol{X}). \qquad \|\boldsymbol{E}\|_1 = \textstyle\sum_{ij} |\boldsymbol{E}_{ij}|.$$



Nuclear norm heuristic: [Fazel, Hindi, Boyd '01], see also [Recht, Fazel, Parillo '08]

MAIN RESULT – Correct recovery

Theorem 1 (Principal Component Pursuit). *If* $\boldsymbol{X}_0 \in \mathbb{R}^{m \times n}$, $m \geq n$ *has rank*

$$r \;\leq\; \rho_r \, \frac{n}{\mu \log^2(m)}$$

and $\boldsymbol{E}_0$ *has Bernoulli support with error probability* $\rho \leq \rho_s^\star$, *then with very high probability*

$$(\boldsymbol{X}_0, \boldsymbol{E}_0) \;=\; \arg\min \|\boldsymbol{X}\|_* + \frac{1}{\sqrt{m}} \|\boldsymbol{E}\|_1 \quad \text{subj} \quad \boldsymbol{X} + \boldsymbol{E} = \boldsymbol{X}_0 + \boldsymbol{E}_0,$$

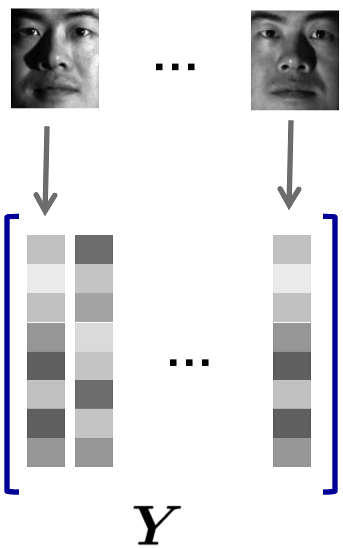
and the minimizer is unique.

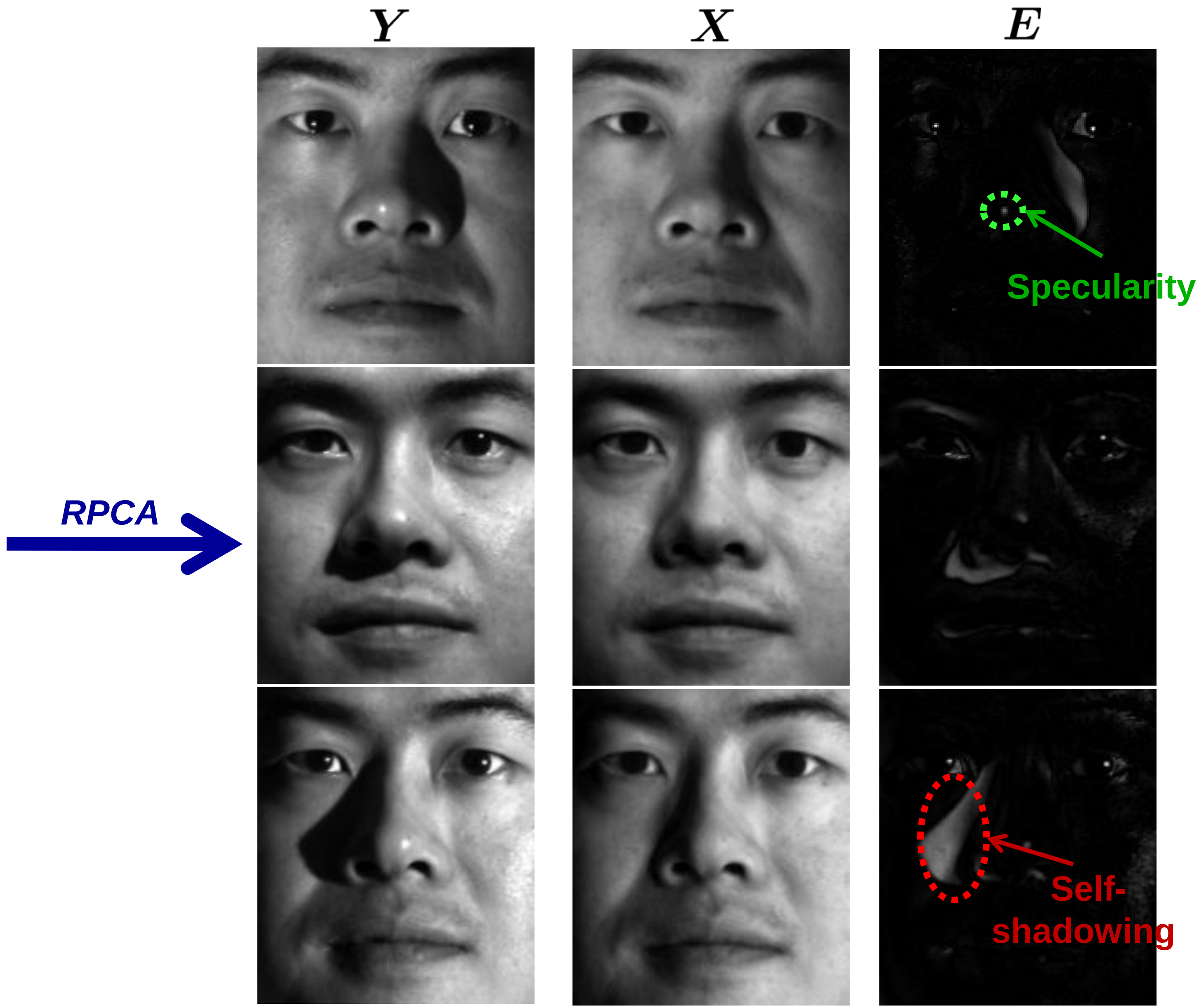
"Convex optimization recovers matrices of rank $O\left(\frac{n}{\log^2 m}\right)$ ***from errors corrupting*** $O\left(mn\right)$ ***entries"***

[Candès, Li, Ma, and W., '09].

EXAMPLE – Faces under varying illumination

58 images of one person under varying lighting:



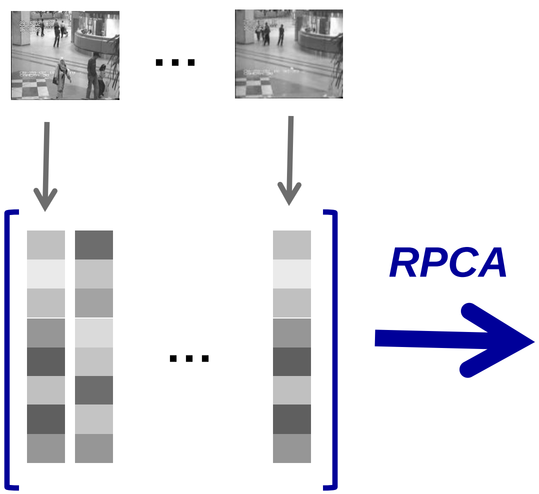


APPLICATIONS – Background modeling from video

Static camera surveillance video

200 frames,
144 x 172 pixels,

Significant foreground motion



$\boldsymbol{Y}$

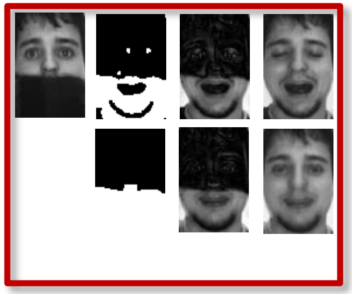
Video Y = Low-rank appx. X + Sparse error E

BIG PICTURE – Parallelism of Sparsity and Low-Rank

	Sparse Vector	*Low-Rank Matrix*
Degeneracy of	one signal	correlated signals
Measure	ℓ^0 norm $\|x\|_0$	$\mathrm{rank}(X)$
Convex Surrogate	ℓ^1 norm $\|x\|_1$	Nuclear norm $\|X\|_*$
Compressed Sensing	$y = Ax$	$Y = A(X)$
Error Correction	$y = Ax + e$	$Y = A(X) + E$
Domain Transform	$y \circ \tau = Ax + e$	$Y \circ \tau = A(X) + E$
Mixed Structures	$Y = A(X) + B(E) + Z$	

A SUITE OF POWERFUL REGULARIZERS …

… for recovering various types of low-dimensional structure:



- [Zhou et. al. '09] Spatially contiguous sparse errors via MRF
- [Bach '10] – structured relaxations from submodular functions
- [Negahban+Yu+Wainwright '10] – geometric analysis of recovery
- [Becker+Candès+Grant '10] – algorithmic templates
- [Xu+Caramanis+Sanghavi '11] – column sparse errors $L_{2,1}$ norm
- [Recht+Parillo+Chandrasekaran+Wilsky '11] – compressive sensing of various structures
- [Candes+Recht '11] – compressive sensing of decomposable structures

$$X^0 = \arg\min \|X\|_\diamond \quad \text{s.t.} \quad \mathcal{P}_Q(X) = \mathcal{P}_Q(X^0)$$

- [McCoy+Tropp'11] – decomposition of sparse and low-rank structures

$$(X_1^0, X_2^0) = \arg\min \|X_1\|_{(1)} + \lambda \|X_2\|_{(2)} \ \text{ s.t. } \ X_1 + X_2 = X_1^0 + X_2^0$$

- [W.+Ganesh+Min+Ma, ISIT'12] – superposition of decomposable structures

$$(X_1^0, \ldots, X_k^0) = \arg\min \sum \lambda_i \|X_i\|_{(i)} \ \text{ s.t. } \ \mathcal{P}_Q(\textstyle\sum_i X_i) = \mathcal{P}_Q(\sum_i X_i^0)$$

Take home message:

Let the data / application tell you the structure…

THANK YOU!

Questions, please?

Next: Efficient, **scalable algorithms** for large ℓ^1 and nuclear norm problems

Later: More **applications** of these techniques

A FEW REFERENCES

General surveys:

Donoho, Elad and Bruckstein, F*rom Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images*, SIAM Review ’09 (see also Elad’s book).

Davenport, Duarte, Eldar, Kityniok, *Introduction to Compressed Sensing*, Signal Processing Magazine ‘11

W., Ma, Sapiro, Mairal, Huang, Yan, *Sparse Representation for Computer Vision and Pattern Recognition*, Proc. IEEE ‘10

Hardness of L0 minimization:

Natarajan, *Sparse Approximate Solutions to Linear Systems*, SIAM Journal on Computing ’95

Amaldi and Kann, *On the Approximability of Minimizing Nonzero variables or Unsatisfied Relations in Linear Systems*, Theoretical Computer Science ’97

Uniqueness of sparse solutions:

Donoho and Elad, *Optimally Sparse Representations in General (nonorthogonal) Dictionaries via L1 Minimization*, PNAS ’03

Gorodnitsky and Rao, *Sparse Signal Reconstruction from Limited Data using FOCUSS – A Reweighted Minimum Norm Algorithm*, IEEE Trans. Signal Processing ’97

The L1 relaxation:

Tibshirani, *Regression shrinkage and selection via the LASSO*, Journal of the Royal Statistical Society Series B, ‘96

Chen, Donoho and Saunders, *Atomic Decomposition by Basis Pursuit*, SIAM Review ’98

(In)-coherence and recovery guarantees:

Donoho and Elad, *Optimally Sparse Representations in General (nonorthogonal) Dictionaries via L1 Minimization*, PNAS ’03

Gribonval and Nielsen, *Sparse Representations in Unions of Bases*, IEEE Info Thy ‘04

Fuchs, *On Sparse Representations in Arbitrary Redundant Bases*, IEEE Info Thy ’04

A FEW MORE REFERENCES

RIP Based Guarantees for L1 Minimization

Candes and Tao, *Near Optimal Signal Recovery from Random Projections: Universal Encoding Strategis?* IEEE Trans Info Thy, ‘04

Candes and Tao, *Decoding by Linear Programming*, IEEE Trans Info Thy, ’05

Candes, *The Restricted Isometry Property and Its Implications for Compressed Sensing*, ‘08

A Few Related Results for Sparse Recovery (not covered in lecture)

Tropp, *Just Relax: Convex Programming Methods for Recovering Sparse Signals in Noise*, IEEE Info Thy ’06

Wainwright, *Sharp Thresholds for Noisy and High-Dimensional Recovery of Sparsity Using L1 Constrained Quadratic Programming (LASSO)*, IEEE Info Thy ‘09

Donoho and Tanner, *Counting Faces of Randomly Projected Polytopes when Projection Radically Lowers Dimension*, Journal of the AMS ‘09

Bayati, Lelarge and Montanari, *Universality of Polytope Phase Transitions and Message Passing Algorithms*, ’12

A Few Applications in this Lecture (more in Lecture 3)

Lustig, Donoho, Santos and Pauly, *Compressive Sensing MRI*, Magnetic Resonance Medicine, ’07

Horev, Bryt and Rubinstein, *Adaptive Image Compression using Sparse Dictionaries,* ’12

W., Yang, Ganesh, Sastry, Ma, *Robust Face Recognition via Sparse Representation*, IEEE PAMI ’09

Elhamifar and Vidal, *Sparse Subspace Clustering*, CVPR ‘09

A FEW *MORE* REFERENCES

Classical Matrix Approximations (PCA)

Pearson, *On lines and planes best fit to systems of points in space*, Philosophical Magazine 1901

Hotelling, *Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology, 1933

Nuclear Norm Minimization

Fazel, Hindi, Boyd, *A Rank Minimization Heuristic with Application to Minimum-Order System Identification*, ACC ‘01

Recht, Fazel, Parillo, *Guaranteed Minimum rank Solutions to Linear Matrix Equations via Nuclear Norm Minimization*, SIAM Review ‘10

Matrix Completion

Candes and Recht, *Exact Matrix Completion via Convex Optimization*, Foundations of Computational Mathematics ‘09

Gross, *Recovering Low-Rank Matrices from Few Coefficients in Any Basis*, IEEE Trans. Info. Thy., ’10

Robust PCA and Matrix Decompositions

Candes, Li, Ma, W., *Robust Principal Component Analysis?* Journal of the ACM ‘11

Chandrasekaran, Sanghavi, Pararilo and Wilsky, *Rank-Sparsity Incoherence for Matrix Decomposition*, SIAM Journal on Optimization, ’11

Agarwal, Negahban and Wainwright, *Noisy Matrix Decomposition via Convex Relaxation: Optimal Rates in High Dimensions* ‘12

This is a huge (and hugely active) area! Many important ideas and papers are missing from the above list. You can complement your reading by visiting …

The UIUC Matrix Recovery site: http://perception.csl.illinois.edu/matrix-rank/home.html

The Rice Compressed Sensing Archive: http://dsp.rice.edu/cs

Nuit Blanche (a blog in this area): http://nuit-blanche.blogspot.com/