Low Rank Matrix Approximations

**Acknowledgement:** The slides are adapted from the ones prepared by Y. Ma and J. Wright.
How to extract **low-dim structures** from such **high-dim data**?
Visual data exhibit **low-dimensional structures** due to rich **local** regularities, **global** symmetries, **repetitive** patterns, or **redundant** sampling.
If we view the data (image) as a matrix

\[ A = [a_1 \mid \cdots \mid a_n] \in \mathbb{R}^{m \times n} \]

then

\[ r \doteq \text{rank}(A) \ll m. \]

**Principal Component Analysis (PCA) via singular value decomposition (SVD):**

- Optimal estimate of $A$ under iid Gaussian noise $D = A + Z$
- Efficient and scalable computation
- Fundamental statistical tool, with huge impact in image processing, vision, web search, bioinformatics…

But… **PCA breaks down under even a single corrupted observation.**
Real application data often contain **missing observations, corruptions**, or subject to unknown **deformation or misalignment**.

**Classical methods** (e.g., PCA, least square regression) **break down**...
Everything old …

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

R. J. Boscovich. De calculo probabilitatum 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
... is new again

Today, robust estimation in high dimension is more urgent and increasingly better understood.

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

**Algorithms** – large scale convex optimization, geometric convergence rate, parallel and distributed computing …

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

Tukey, Bickel, Huber, Hampel, Tibishirani, Donoho, … Candes and Tao 2004 …

and many more I will mention later…
The data should be **low-dimensional**:

\[ A = [a_1 \mid \cdots \mid a_n] \in \mathbb{R}^{m \times n}, \quad \text{rank}(A) \ll m. \]
The data should be **low-dimensional**: 

\[ A = [a_1 | \cdots | a_n] \in \mathbb{R}^{m \times n}, \quad \text{rank}(A) \ll m. \]

…but some of the observations are **grossly corrupted**:

\[ A + E, \quad |E_{ij}| \]

\( E_{ij} \) arbitrarily large, but most are zero.
The data should be low-dimensional:

\[ A = [a_1 | \cdots | a_n] \in \mathbb{R}^{m \times n}, \quad \text{rank}(A) \ll m. \]

… but some of the observations are grossly corrupted:

\[ A + E, \quad |E_{ij}| \]

\( E_{ij} \) arbitrarily large, but most are zero.

… and some of them can be missing too:

\[ D = \mathcal{P}_\Omega[A + E], \]

\( \Omega \subset [m] \times [n] \) the set of observed entries.
The data should be **low-dimensional**:

\[ A = [a_1 | \cdots | a_n] \in \mathbb{R}^{m \times n}, \quad \text{rank}(A) \ll m. \]

… but some of the observations are **grossly corrupted**:

\[ A + E, \quad |E_{ij}| \]

\[ E_{ij} \text{ arbitrarily large, but most are zero.} \]

… and some of them can be **missing** too:

\[ D = \mathcal{P}_\Omega[A + E], \]

\[ \Omega \subset [m] \times [n] \text{ the set of observed entries.} \]

… special cases of a more general problem:

\[ D = \mathcal{L}_1(A) + \mathcal{L}_2(E) + Z \quad A, E \text{ either sparse or low-rank} \]
Our problem:

Given an observation \( D = \mathcal{P}_\Omega [A + E + Z] \), with

\[
\begin{align*}
A & \; \text{low-rank}, \\
E & \; \text{sparse}, \\
Z & \; \text{small, dense noise},
\end{align*}
\]

recover a good estimate of \( A \).

High-dimensional setting: Many more unknowns \((A, E)\) than observations \(\Omega\). Can knowledge of the structure of \((A, E)\) make this problem well-posed?

Hugely active area...
Our problem:

Given an observation \( D = \mathcal{P}_\Omega[A + E + Z] \), with

- \( A \): low rank,
- \( E \): sparse,
- \( Z \): small, dense noise,

recover a good estimate of \( A \).

**High-dimensional setting:** Many more unknowns \( (A, E) \) than observations \( \Omega \). Can knowledge of the structure of \((A, E)\) make this problem well-posed?

**Hugely active area...**
**CONTEXT – Recent related progress**

**Sparse recovery:** Given \( y = Lx_0, \) \( L \in \mathbb{R}^{m \times n}, \) \( m \ll n, \) recover \( x_0.\)

\[
\begin{align*}
\begin{bmatrix}
\vdots \\
\vdots \\
\vdots \\
\end{bmatrix}
&= \\
\begin{bmatrix}
\vdots \\
\vdots \\
\vdots \\
\end{bmatrix} \\
\begin{bmatrix}
\vdots \\
\vdots \\
\vdots \\
\end{bmatrix}
&= \\
\begin{bmatrix}
\vdots \\
\vdots \\
\vdots \\
\end{bmatrix}
\end{align*}
\]

**Impossible** in general \( (m \ll n) \)

**Well-posed** if \( x_0 \) is structured (sparse), but still **NP-hard**

**Tractable** via convex optimization: \( \min \|x\|_1 \) s.t. \( y = Lx \)

... if \( L \) is “nice” (random, incoherent, RIP)

**Hugely active area:** Donoho+Huo ’01, Elad+Bruckstein ’03, Candès+Tao ’04,’05, Tropp ’04, ’06, Donoho ’04, Fuchs ’05, Zhao+Yu ’06, Meinshausen+Buhlmann ’06, Wainwright ’09, Donoho +Tanner ’09 ... and many others
Our problem:

Given an observation $D = P_\Omega[A + E + Z]$, with

- $A$ low-rank,
- $E$ sparse,
- $Z$ small, dense noise,

recover a good estimate of $A$.

High-dimensional setting: Many more unknowns $(A, E)$ than observations $\Omega$. Can knowledge of the structure of $(A, E)$ make this problem well-posed?

Hugely active area...
Our problem:

Given an observation \( D = \mathcal{P}_\Omega[A + E + Z] \), with

- \( A \) low-rank,
- \( E \) sparse,
- \( Z \) small, dense noise,

recover a good estimate of \( A \).

High-dimensional setting: Many more unknowns \((A, E)\) than observations \(\Omega\). Can knowledge of the structure of \((A, E)\) make this problem well-posed?

Hugely active area...
Low-rank sensing: Given $y = \mathcal{L}[A_0]$, $\mathcal{L} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^p$, recover $A_0$.

- Impossible in general ($p \ll mn$)
- Well-posed if $A_0$ is structured (low-rank), but still NP-hard
- Tractable via convex optimization: $\min \|A\|_* \text{ s.t. } y = \mathcal{L}(A)$
  ... if $\mathcal{L}$ is “nice” (random, rank-RIP)

Hugely active area: Recht+Fazel+Parillo ’07, Candès+Plan ’10, Mohan+Fazel ‘10, Recht +Xu+Hassibi ’11, Chandrasekaran+Recht+Parillo+Willsky ’11, Negahban+Wainwright ’11 …
Our problem:

Given an observation $D = \mathcal{P}_\Omega[A + E + Z]$, with

- $A$ low-rank,
- $E$ sparse,
- $Z$ small, dense noise,

recover a good estimate of $A$.

**High-dimensional setting:** Many more unknowns $(A, E)$ than observations $\Omega$. Can knowledge of the structure of $(A, E)$ make this problem well-posed?

**Hugely active area...**
Our problem:

Given an observation \( D = \mathcal{P}_\Omega [A + E + Z] \), with

- \( A \) low-rank,
- \( E \) sparse,
- \( Z \) small, dense noise,

recover a good estimate of \( A \).

**High-dimensional setting:** Many more unknowns \((A, E)\) than observations \(\Omega\). Can knowledge of the structure of \((A, E)\) make this problem well-posed?

**Hugely active area...**
Matrix completion: Given \( y = \mathcal{P}_\Omega [A_0], \Omega \subset [m] \times [n], \) recover \( A_0. \)

**Impossible** in general (\( |\Omega| \ll mn \))

**Well-posed** if \( A_0 \) is structured (low-rank), but still **NP-hard**

**Tractable** via convex optimization: 
\[
\min \| A \|_* \text{ s.t. } y = \mathcal{P}_Q (A)
\]

... if \( \Omega \) is “nice” (random subset) ...

... and \( A_0 \) interacts “nicely” with \( \mathcal{P}_\Omega (A_0 \text { incoherent – not “spiky”}). \)

**Hugely active area:** Candès+Recht ‘08, Keshevan+Oh+Montonari ‘09, Candès+Tao ‘09, Gross ‘10, Recht ‘10, Negahban+Wainwright ‘10
Our problem:

Given an observation $D = \mathcal{P}_\Omega [A + E + Z]$, with

- $A$ low-rank,
- $E$ sparse,
- $Z$ small, dense noise,

recover a good estimate of $A$.

High-dimensional setting: Many more unknowns $(A, E)$ than observations $\Omega$. Can knowledge of the structure of $(A, E)$ make this problem well-posed?

Hugely active area...
Our problem: Given an observation $D = \mathcal{P}_\Omega [A + E + Z]$, with

- $A$ low-rank,
- $E$ sparse,
- $Z$ small, dense noise,

recover a good estimate of $A$.

High-dimensional setting: Many more unknowns $(A, E)$ than observations $\Omega$. Can knowledge of the structure of $(A, E)$ make this problem well-posed?

Hugely active area...
Robust recovery: Given $y = Lx_0 + e_0$, $L \in \mathbb{R}^{m \times n}$, $m \ll n$, recover $x_0$ and $e_0$.

Impossible in general ($m \ll n + m$)

Well-posed if $x_0$ is sparse, errors $e_0$ not too dense, but still **NP-hard**

**Tractable:** via convex optimization: $\min \|x\|_1 + \|e\|_1$ s.t. $y = Lx + e$

… if $L$ is “nice” (cross and bouquet)

**Hugely active area:** Candès+Tao ’05, Wright+Ma ’10, Nguyen+Tran ‘11, Li ’11, also Zhang, Yang, Huang’11, etc…
Given an observation \( D = \mathcal{P}_\Omega[A + E + Z] \), with

- \( A \) low-rank,
- \( E \) sparse,
- \( Z \) small, dense noise,

recover a good estimate of \( A \).

- Theory and Algorithm
- Potential Applications
Our problem:

Given an observation $D = \mathcal{P}_\Omega[A + E + Z]$, with

- $A$ low-rank,
- $E$ sparse,
- $Z$ small, dense noise,

recover a good estimate of $A$.

Numerous approaches in the literature:

- Random sampling [Fischler and Bolles ‘81]
- Multivariate trimming [Gnanadesikan and Kettering ‘72]
- Power Factorization [Wieber’70s, Shum & Ikeuchi’96]
- Alternating minimization [Ke and Kanade ‘03]
- Influence functions [de la Torre and Black ‘03]

Key concern: can we guarantee correctness with an efficient algorithm?
The problem is hard: formally

$$D = \mathcal{P}_\Omega[ A + E + Z ]$$

low-rank sparse

**Matrix rigidity** [Valiant ‘77]: How many entries do we need to change to decrease the rank?

$$\mathcal{R}_D(r) = \min \| E \|_0 \quad \text{subject to} \quad A + E = D, \ \text{rank}(A) \leq r.$$ 

Related to questions in computational complexity (fast evaluation of linear maps), communication complexity, etc.

Computing rigidity over finite fields is NP-hard [Deshpande ‘07].
The problem is hard: intuitively

\[ D = \mathcal{P}_\Omega[ A + E + Z ] \]

low-rank sparse

Some very sparse matrices are also low-rank:

\[ D = 1_{ij} \]

\[ A = 1_{ij} \quad E = 0 \]

\[ A = 0 \quad E = 1_{ij} \]

Ambiguity if the low rank term \( A \) “looks sparse”.
When is there hope?

\[ D = \mathcal{P}_\Omega [ A + E + Z ] \]

Can hope for success when \( A \) is not too sparse. Formally, when the singular vectors of \( A = U S V^* \) are incoherent:

- Singular vectors are not too spiky:
  \[
  \max_i \| U_i \|^2 \leq \mu r / m.
  \]
  \[
  \max_i \| V_i \|^2 \leq \mu r / n.
  \]

- … and not cross-correlated:
  \[
  \| U V^* \|_\infty \leq \sqrt{\mu r / mn}
  \]

Incoherence condition: [Candès+Recht ‘08].

First condition bounds the maximum leverage score – used in robust regression [McCoy+Tropp ‘11], approximate linear algebra [Mahoney ‘11].
The problem is hard: intuitively

\[ D = P_\Omega[ A + E + Z ] \]

low-rank sparse

Certain sparse error patterns \( E \) make recovering \( A \) impossible:

\[
\begin{bmatrix}
A
\end{bmatrix}
+ \begin{bmatrix}
E = e_i v^T
\end{bmatrix}
= \begin{bmatrix}
D
\end{bmatrix}
\]

Here, cannot hope to recover the second row of \( A \). Ambiguity when support of \( E \) is adversarial…
When is there hope?

\[ D = \mathcal{P}_\Omega[A + E + Z] \]

low-rank sparse

Can hope for success when \( E \) is not too concentrated on any row or column. Formalize this via a random error support model:

Uniform model on error support, signs and magnitudes arbitrary:

\[
\text{supp}(E) \sim \text{uni}(m \times n) \quad \Omega \sim \text{uni}(m \times n)
\]

Other models, e.g. Markov Random Field, also possible.
Disambiguating the solution...

\[ D = \mathcal{P}_\Omega [ A + E + Z ] \]

\text{low-rank} \quad \text{sparse}

Singular vectors of \( A \) are \textit{incoherent}, singular values \textit{arbitrary}.

Error \( E \) has random support, \textit{signs and magnitudes arbitrary}.

Noise is not too large: \( \| Z \|_F \leq \eta \).

\textbf{Under these assumptions, problem is well-posed.}
... and how might we solve it?

\[ D = \mathcal{P}_\Omega [ A + E + Z ] \]

**Naïve optimization approach**

Look for a low-rank \( A \) that agrees with the data up to some sparse error \( E \):

\[
\min \ \text{rank}(A) + \gamma \|E\|_0 \ \text{subj} \ \mathcal{P}_\Omega [A + E] = D.
\]

\[
\text{rank}(A) = \# \{ \sigma_i(A) \neq 0 \}. \quad \|E\|_0 = \# \{ E_{ij} \neq 0 \}.
\]
... and how might we solve it?

\[ D = \mathcal{P}_\Omega[A + E + Z] \]

**Naive optimization approach**

Look for a low-rank \( A \) that agrees with the data up to some sparse error \( E \):

\[
\begin{align*}
\min & \quad \text{rank}(A) + \gamma \|E\|_0 \\
\text{subj} & \quad \mathcal{P}_\Omega[A + E] = D.
\end{align*}
\]

\[
\text{rank}(A) = \#\{\sigma_i(A) \neq 0\}. \quad \|E\|_0 = \#\{E_{ij} \neq 0\}.
\]

**INTRACTABLE**
... and how might we solve it?

\[ D = \mathcal{P}_\Omega[ A + E + Z ] \]

**Naïve optimization approach**

Look for a low-rank \( A \) that agrees with the data up to some sparse error \( E \):

\[
\begin{align*}
\min \ & \text{rank}(A) + \gamma \| E \|_0 \\
\text{subj} \ & \mathcal{P}_\Omega[A + E] = D.
\end{align*}
\]

**Convex relaxation**

\[
\begin{align*}
\text{rank}(A) &= \# \{ \sigma_i(A) \neq 0 \}. \\
\| E \|_0 &= \# \{ E_{ij} \neq 0 \}. \\
\| A \|_* &= \sum_i \sigma_i(A). \\
\| E \|_1 &= \sum_{ij} |E_{ij}|.
\end{align*}
\]

Nuclear norm heuristic: Fazel, Hindi, Boyd '01; Recht, Fazel, Parillo '08.
... and how might we solve it?

\[ D = \mathcal{P}_\Omega[ A + E + Z ] \]

**Naïve optimization approach**

Look for a low-rank \( A \) that agrees with the data up to some sparse error \( E \):

\[
\begin{align*}
\min \ & \text{rank}(A) + \gamma \|E\|_0 \\
\text{s.t.} \ & \mathcal{P}_\Omega[A + E] = D.
\end{align*}
\]

**Convex relaxation**

\[
\begin{align*}
\min \ & \|A\|_* + \lambda \|E\|_1 \\
\text{s.t.} \ & \mathcal{P}_\Omega[A + E] = D.
\end{align*}
\]

Semidefinite program, solvable in polynomial time – “efficient” algorithm.  
**Practical** thanks to very recent advances – second talk of this tutorial.
**KEY QUESTION**

\[ D = \mathcal{P}_\Omega[ A + E + Z ] \]

**Key question**

Does this practical surrogate actually solve the problem?

\[
\min \; \text{rank}(A) + \gamma \| E \|_0 \quad \text{subj} \quad \mathcal{P}_\Omega[A + E] = D.
\]

\[
\min \; \| A \|_* + \lambda \| E \|_1 \quad \text{subj} \quad \mathcal{P}_\Omega[A + E] = D.
\]

**Not always** – original problem is NP-hard.

But maybe it succeeds for the cases we care about?
Does this actually work?

(a) Robust PCA, Random Signs  
(b) Robust PCA, Coherent Signs  
(c) Matrix Completion

\[ D = A + E \]  
\[ D = P_{\Omega}[A] \]

**Apparently yes...** white regions are problems with perfect recovery.

Correct recovery when \( A \) is indeed **low-rank** and \( E \) is indeed **sparse**?
Does this actually work?

\[ D = A + E \]

Apparently yes... it does interesting things on real data (more later on).
MAIN RESULT I: exact solution in noise-free case

\[ D = \mathcal{P}_\Omega[ A_0 + E_0 ], \quad Z = 0. \]

Theorem 1. If \( A_0, E_0 \in \mathbb{R}^{m \times n}, m \geq n, \) with

\[
\text{rank}(A_0) \leq C \frac{n}{\mu \log^2(m)}, \quad \text{and} \quad \|E_0\|_0 \leq \rho^* mn,
\]

and we observe only a random subset of size

\[ |\Omega| = mn/10 \]

entries, then with very high probability, solving the convex program

\[
\min \|A\|_* + \frac{1}{\sqrt{m}} \|E\|_1 \quad \text{subj} \quad P_\Omega[A + E] = D,
\]

uniquely recovers \((A_0, E_0)\).
Theorem 1. If $A_0, E_0 \in \mathbb{R}^{m \times n}, m \geq n$, with a very high probability, solving the convex program

\[
\min \|A\|_* + \frac{1}{\sqrt{m}} \|E\|_1 \quad \text{subj} \quad P_\Omega[A + E] = D,
\]

uniquely recovers $(A_0, E_0)$.

“Convex optimization exactly recovers matrices of rank $O\left(\frac{n}{\log^2 m}\right)$ from $O\left(mn\right)$ errors, with large fractions of the observations missing!”
MAIN RESULT I: exact solution in noise-free case

\[ D = \mathcal{P}_\Omega [ A_0 + E_0 ], \quad Z = 0. \]

Theorem 1. If \( A_0, E_0 \in \mathbb{R}^{m \times n}, m \geq n, \) with

\[
\text{rank}(A_0) \leq C \frac{n}{\mu \log^2(m)}, \quad \text{and} \quad \|E_0\|_0 \leq \rho^* mn,
\]

and we observe only a random subset of size \textbf{Non-adaptive weight factor} entries, then with very high probability, solving the convex program

\[
\min \|A\|_* + \frac{1}{\sqrt{m}} \|E\|_1 \quad \text{subj} \quad P_\Omega[A + E] = D,
\]

uniquely recovers \((A_0, E_0)\).
MAIN RESULT II: stability under noise

\[ D = \mathcal{P}_{\Omega}[ A_0 + E_0 + Z ], \quad \|Z\|_F \leq \eta. \]

**Theorem 2.** If \( A_0, E_0 \) satisfy the conditions of the previous result (i.e., exact recovery occurs in the noise free case), then in the noisy case, the recovery is stable. Formally,

\[
(A, E) = \arg\min \|A\|_* + \frac{1}{\sqrt{m}} \|E\|_1 \quad \text{subj \quad } \|\mathcal{P}_\Omega[A + E] - D\|_F \leq \eta,
\]

then

\[
\|(A, E) - (A_0, E_0)\|_F \leq C\eta.
\]
MAIN RESULT II: stability under noise

\[ D = \mathcal{P}_\Omega[ A_0 + E_0 + Z ], \quad \| Z \|_F \leq \eta. \]

Theorem 2. If \( A_0, E_0 \) satisfy the conditions of the previous result (i.e., exact recovery occurs in the noise free case), then in the noisy case, the recovery is stable.

“Convex optimization recovers matrices of from \( O\left(\frac{n}{\log^2 m}\right) \) with large fractions of the \( O\left(mn\right) \) observations missing, and the estimation error is proportional to the noise in the uncorrupted entries!”

Zhou, Li, W., Candès, Ma, ISIT 2010.
MAIN RESULT III: dense error correction

\[ D = A_0 + E_0, \quad Z = 0. \]

Theorem 3 (Dense Error Correction). If \( A_0 \) has rank \( r \leq \rho r \mu^2 \frac{m}{\log^2(n)} \) and \( E_0 \) has random signs and Bernoulli support with error probability \( \rho < 1 \), then with very high probability

\[(A_0, E_0) = \arg \min \| A \|_* + \lambda \| E \|_1 \quad \text{subj} \quad A + E = A_0 + E_0,\]

and the minimizer is unique.

“Convex optimization corrects nearly 100% corruptions!”
**BIG PICTURE – Landscape of Theoretical Guarantees**

What people have done and known so far in the past 3-4 years:

**Matrix Recovery (RPCA)**

- rank \(=O\left(\frac{m}{\log^2 n}\right)\)
- \(\|E_0\|_0\) vs. \(mn\)
- \(\text{rank}(A_0)\) vs. \(m\)

**Matrix Completion**

- rank \(=O\left(\frac{m}{\log^2 n}\right)\)
- \(\|E_0\|_0\) vs. \(mn\)
- \(\|E_0\|_0 < (m - 6r)m\)
ALGORITHMS – Are scalable solutions possible?

Seemingly BAD NEWS: Our optimization problem

$$\min_{A,E} \|A\|_1 + \lambda \|E\|_1 \quad \text{s.t. } A + E = D$$

is high-dimensional and non-smooth.

Convergence rate of solving a generic convex program:

$$\min_{x} f(x)$$

# of iterations Second-order Newton method is unscalable!
First-order methods depend strongly on the smoothness of \( f \)!
ALGORITHMS – Are scalable solutions possible?

GOOD NEWS: The objective function has special structures

$$\min_{A;E} \left\| A \right\|_* + \lambda \left\| E \right\|_1 \quad \text{s.t. } A + E = D$$

KEY OBSERVATION: closed form solutions for the proximal minimizations can be derived!

Solutions are given by soft-thresholding the entries and singular values of the matrix, respectively:
ALGORITHMS – *Evolution of scalable algorithms*

**GOOD NEWS:** Scalable first-order gradient-descent algorithms:
- Iterative Thresholding [Osher, Mao, Dong, Yin ’09, Wright et. al.’09, Cai et. al.’09].
- Accelerated Proximal Gradient [Nesterov ’83, Beck and Teboulle ’09]:
- Augmented Lagrange Multiplier [Hestenes ‘69, Powell ’69]:
- Alternating Direction Method of Multipliers [Gabay and Mercier ’76].

**A scalable algorithm:** alternating direction method (ADM) for ALM:

\[
l(A, E, Y) = \|A\|_* + \lambda \|E\|_1 + \langle Y, D - A - E \rangle + \frac{\mu}{2} \|D - A - E\|_F^2
\]

\[
A_{k+1} = D_{\mu_k^{-1}}(D - E_k + Y_k/\mu_k), \quad \text{Shrink singular values}
\]

\[
E_{k+1} = S_{\lambda \mu_k^{-1}}(D - A_{k+1} + Y_k/\mu_k), \quad \text{Shrink absolute values}
\]

\[
Y_{k+1} = Y_k + \mu_k(D - A_{k+1} - E_{k+1}).
\]

Cost of each iteration is a classical PCA, i.e. a (partial) SVD.

Lin, Chen, and Ma, UILU-ENG-09-2214, 2010.
ALGORITHMS – *Evolution of fast algorithms (around 2009)*

For a 1000x1000 matrix of rank 50, with 10% (100,000) entries randomly corrupted:

\[
\min \| A \|_* + \lambda \| E \|_1 \quad \text{subj} \quad A + E = D.
\]

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Accuracy</th>
<th>Rank</th>
<th>|E|₀</th>
<th># iterations</th>
<th>time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>5.99e-006</td>
<td>50</td>
<td>101,268</td>
<td>8,550</td>
<td>119,370.3</td>
</tr>
<tr>
<td>DUAL</td>
<td>8.65e-006</td>
<td>50</td>
<td>100,024</td>
<td>822</td>
<td>1,855.4</td>
</tr>
<tr>
<td>APG</td>
<td>5.85e-006</td>
<td>50</td>
<td>100,347</td>
<td>134</td>
<td>1,468.9</td>
</tr>
<tr>
<td>APGₚ</td>
<td>5.91e-006</td>
<td>50</td>
<td>100,347</td>
<td>134</td>
<td>82.7</td>
</tr>
<tr>
<td>EALMₚ</td>
<td>2.07e-007</td>
<td>50</td>
<td>100,014</td>
<td>34</td>
<td>37.5</td>
</tr>
<tr>
<td>IALMₚ</td>
<td>3.83e-007</td>
<td>50</td>
<td>99,996</td>
<td>23</td>
<td>11.8</td>
</tr>
</tbody>
</table>

*Provably Robust PCA at only a constant factor (≈20) more computation than conventional PCA!*
ALGORITHMS – *Convergence rate with strong convexity*

**GREAT NEWS:** Geometric convergence for gradient algorithms!

[Agarwal, Negahban, Wainwright, NIPS 2010]

---

Figure 1. Convergence rates of projected gradient descent in application to Lasso programs ($\ell_1$-constrained least-squares). Each panel shows the log optimization error $\log \|\hat{\theta}^t - \hat{\theta}\|$ versus the iteration number $t$. Panel (a) shows three curves, corresponding to dimensions $d \in \{5000, 10000, 20000\}$, sparsity $s = [\sqrt{d}]$, and all with the same sample size $n = 2500$. All cases show geometric convergence, but the rate for larger problems becomes progressively slower. (b) For an appropriately rescaled sample size ($\alpha = \frac{n}{\log d}$), all three convergence rates should be roughly the same, as predicted by the theory.
APPLICATIONS

- Repairing Images and Videos
  - Image Repairing, Background Extraction, Street Panorama

- Reconstructing 3D Geometry
  - Shape from Texture, Featureless 3D Reconstruction

- Registering Multiple Images
  - Multiple Image Alignment, Video Stabilization

- Recognizing Objects
  - Faces, Texts, etc

- Other Data and Applications
Implications: Highly Compressive Sensing of Structured Information!

*Recover low-dimensional structures from a fraction of missing measurements with structured support.*

compressive samples ➞ Low-rank Structures + Sparse Structures
Repairing Images: Highly Robust Repairing of Low-rank Textures!

Liang, Ren, Zhang, and Ma, in ECCV 2012.
Repairing Low-rank Textures

Low-rank Method vs Photoshop

Input

Output
Repairing (Distorted) Low-rank Textures

Low-rank Method

Photoshop

Input

Output
Structured Texture Completion and Repairing
Repairing Multiple Correlated Images

58 images of one person under varying lighting:

\[ D \]

\[ A \]

\[ E \]

specularity

cast shadows

Repairing Images: robust photometric stereo

\[
\min \| A \|_* + \lambda \| E \|_1 \quad \text{subj} \quad D = \mathcal{P}_\Omega (A + E).
\]

\[\Omega^c \sim \text{shadow}(20.7\%) \quad E \sim \text{specularities}(13.6\%)
\]

Mean error  
Max error

\[
\begin{align*}
\text{Wu, Ganesh, Li, Matsushita, and Ma, in ACCV 2010.}
\end{align*}
\]
Repairing Video Frames: \textit{background modeling from video}

Surveillance video

200 frames,
144 x 172 pixels,

Significant foreground motion

\text{Video} = \text{Low-rank appx.} \ A^+ \ + \ \text{Sparse error} \ E

Candès, Li, Ma, and W., JACM, May 2011.
Implications: Highly Compressive Sensing of Structured Information!

*Recover low-dimensional structures from diminishing fraction of corrupted measurements.*
Repairing Video Frames: *Street Panorama*

Zhou, Min, and Ma, submitted to NIPS 2012.
Repairing Video Frames: Street Panorama

Low-rank

AutoStitch

Photoshop
Repairing Video Frames: Street Panorama

Low-rank

AutoStitch

Photoshop
Sensing or Imaging of Low-rank and Sparse Structures

Fundamental Problem: How to recover low-rank and sparse structures from corrupted data subject to either nonlinear deformation $\tau$ or linear compressive sampling $P$?
Reconstructing 3D Geometry and Structures

\[ D \text{ – deformed observation} \]

\[ \circ \tau = A_0 + E_0 \]

\[ A \text{ – low-rank structures} \]

\[ E \text{ – sparse errors} \]

\[ + \]

**Problem:** Given \( D \circ \tau = A_0 + E_0 \), recover \( \tau, A_0 \) and \( E_0 \) simultaneously.

- **Low-rank component** (regular patterns...)
- **Sparse component** (occlusion, corruption, foreground...)

**Parametric deformations** (affine, projective, radial distortion, 3D shape...)

\[ \]
Transform Invariant Low-rank Textures (TILT)

\( D \) – deformed observation \( A \) – low-rank structures \( E \) – sparse errors

\[ \circ \mathcal{T} = \]

\textbf{Objective:} \quad \textit{Principal Component Pursuit:}

\[
\min \quad \|A\|_* + \lambda \|E\|_1 \quad \text{subj} \quad A + E = D \circ \mathcal{T}
\]

\textbf{Solution: Iteratively solving the linearized convex program:}

\[
\min \quad \|A\|_* + \lambda \|E\|_1 \quad \text{subj} \quad A + E = D \circ \tau_k + J \cdot \Delta \mathcal{T}
\]

Or reduced version:

\[
\text{subj} \quad \mathcal{P}_Q[A + E] = \mathcal{P}_Q[D \circ \tau_k], \quad \mathcal{P}_Q[J] = 0
\]

Zhang, Liang, Ganesh, Ma, ACCV’10, IJCV’12
TILT: Shape from texture

Input (red window $D$)

Output (rectified green window $A$)

Zhang, Liang, Ganesh, Ma, ACCV’10, IJCV’12
TILT: Shape and geometry from textures

\[ z = f(x) : x(x - x_m) \sum_{i=1}^{d} a_i x^i \]

\[ D \circ \tau = A + E \]
\[ \tau = (K, R, T, \{a_i\}) \]
TILT: Shape and geometry from textures

360° panorama

Zhang, Liang, and Ma, in ICCV 2011
TILT: Virtual reality

Zhang, Liang, and Ma, in ICCV 2011
TILT: Camera Calibration with Radial Distortion

\[ r = \sqrt{x_0^2 + y_0^2}, f(r) = 1 + kc(1)r^2 + kc(2)r^4 + kc(5)r^6 \]

\[
\begin{pmatrix}
    x \\
    y
\end{pmatrix} = \begin{pmatrix}
    f(r)x_0 + 2kc(3)x_0y_0 + kc(4)(r^2 + 2x_0^2) \\
    f(r)y_0 + 2kc(4)x_0y_0 + kc(3)(r^2 + 2y_0^2)
\end{pmatrix}
\]

\[
K = \begin{bmatrix}
    f_x & 0 & o_x \\
    0 & f_y & o_y \\
    0 & 0 & 1
\end{bmatrix}
\]
TILT: Camera Calibration with Radial Distortion

\[
\min \sum_{i=1}^{N} \| A_i \|_* + \lambda \| E_i \|_1 \quad \text{subj} \quad A_i + E_i = D \circ (\tau_0, \tau_i)
\]

\[
\tau_0 = (K, K_c), \quad \tau_i = (R_i, T_i).
\]

Previous approach

Low-rank method
TILT: Holistic 3D Reconstruction of Urban Scenes

\[
\min \| A \|_* + \| E \|_1 \quad \text{s.t.} \quad A + E = [D_1 \circ \tau_1, D_2 \circ \tau_2]
\]
TILT: Holistic 3D Reconstruction of Urban Scenes

From one input image

From four input images

Mobahi, Zhou, and Ma, in ICCV 2011
**TILT: Holistic 3D Reconstruction of Urban Scenes**

From eight input images

Mobahi, Zhou, and Ma, in ICCV 2011
Virtual reality in urban scenes
Registering Multiple Images: Robust Alignment

\( D – \) corrupted & misaligned observation
\( A – \) aligned low-rank signals
\( E – \) sparse errors

\[ \circ \tau = A_0 + E_0 \]

**Problem:** Given \( D \circ \tau = A_0 + E_0 \), recover \( \tau, A_0 \) and \( E_0 \).

**Solution:** Robust Alignment via Low-rank and Sparse (RASL) Decomposition

**Iteratively solving the linearized convex program:**

\[
\min \| A \|_* + \lambda \| E \|_1 \quad \text{subj} \quad A + E = D \circ \tau_k + J \Delta \tau \\
\text{(or} \quad Q(A + E) = QD \circ \tau_k, QJ = 0)\]
RASL: Aligning Face Images from the Internet

*48 images collected from internet

Peng, Ganesh, Wright, Ma, CVPR’10, TPAMI’11
RASL: Faces Detected

**Input:** faces detected by a face detector ($D$)
RASL: Faces Aligned

Output: aligned faces \((D \circ \tau)\)

Peng, Ganesh, Wright, Ma, CVPR’10, TPAMI’11
RASL: Faces Repaired and Cleaned

Output: clean low-rank faces \((A)\)

Peng, Ganesh, Wright, Ma, CVPR’10, TPAMI’11
RASL: Sparse Errors of the Face Images

Output: sparse error images ($E$)
**RASL: Video Stabilization and Enhancement**

Original video (\(D\))  Aligned video (\(D \circ \tau\))  Low-rank part (\(A\))  Sparse part (\(E\))

Peng, Ganesh, Wright, Ma, CVPR’10, TPAMI’11
RASL: Aligning Handwritten Digits

\[ D \]

\[ D \circ \tau \]

\[ A \]

\[ E \]

Learned-Miller PAMI’06

Vedaldi CVPR’08

Peng, Ganesh, Wright, Ma, CVPR’10, TPAMI’11
Object Recognition: Rectifying Pose of Objects

Input (red window $D$)

Output (rectified green window $A$)

Zhang, Liang, Ganesh, Ma, ACCV’10 and IJCV’12
Object Recognition: Regularity of Texts at All Scales!

Input (red window $D$)

Output (rectified green window $A$)

Zhang, Liang, Ganesh, Ma, ACCV’10 and IJCV’12
Recognition: \textit{Character/Text Rectification}

\[ D \quad D \circ \mathcal{T} = A + E \]

\[ \begin{align*}
\text{TILT} & \\
\text{DCT} & \\
\text{TV} & \\
\end{align*} \quad \begin{align*}
\text{rank-obj} & \\
\text{DCT-obj} & \\
\text{TV-obj} & \\
\end{align*} \]

Xin Zhang, Zhouchen Lin, and Ma, submitted to PR 2012
Recognition: Character/Text Rectification

TILT versus Hough Transform

Xin Zhang, Zhouchen Lin, and Ma, submitted to PAMI 2011
Recognition: Street Sign Rectification

\[ \min \sum_{i=1}^{4} \|A_i\|_1 + \lambda \|E_i\|_1 \]

subj \quad D \circ \tau = [A_1 \cdots A_4] + [E_1 \cdots E_4].

Xin Zhang, Zhouchen Lin, and Ma, submitted to PR 2012
Recognition: Character Rectification and Recognition

Microsoft OCR for rotated characters
(2,500 common Chinese characters)

Microsoft OCR for skewed characters
(2,500 common Chinese characters)

Xin Zhang, Zhouchen Lin, and Ma, submitted to PR 2012
Take-home Messages for Visual Data Processing:

1. (Transformed) **low-rank and sparse** structures are central to visual data modeling, processing, and analyzing;

2. Such structures can now be extracted **correctly, robustly, and efficiently**, from raw image pixels (or high-dim features);

3. These new algorithms **unleash tremendous local or global information** from single or multiple images, emulating or surpassing human capability;

4. These algorithms start to exert significant impact on **image/video processing, 3D reconstruction, and object recognition**.

... ...

*But try not to abuse or misuse them...*
Other Data/Applications: Web Image/Tag Refinement

PROBLEM
Input: images with user-provided tags

SOLUTION
Tag Refinement
Output: images with refined tags

\[ D \rightarrow A + E \]

User-provided tag matrix

\[ A \]
Low-rank matrix

\[ E \]
Sparse error matrix

Content consistency

tag_Animal
tag_Dog

fly bird cool insect strong

fly bird sky eagle

Zhu, Yan, and Ma, ACM MM 2010.
Other Data/Applications: Web Document Corpus Analysis

Latent Semantic Indexing: the classical solution (PCA)

\[ D = A + Z = U_1 \Sigma_1 V_1^T + U_2 \Sigma_2 V_2^T \]

Dense, difficult to interpret

a better model/solution?

\[ D = A + E \]

Low-rank
“background”
topic model
Other Data/Applications: Sparse Keywords Extracted

Reuters-21578 dataset: 1,000 longest documents; 3,000 most frequent words

**CHRYSLER SETS STOCK SPLIT, HIGHER DIVIDEND**

Chrysler Corp said its board declared a three-for-two stock split in the form of a 50 pct stock dividend and raised the quarterly dividend by seven pct.

The company said the dividend was raised to 37.5 cts a share from 35 cts on a pre-split basis, equal to a 25 ct dividend on a post-split basis.

Chrysler said the stock dividend is payable April 13 to holders of record March 23 while the cash dividend is payable April 15 to holders of record March 23. It said cash will be paid in lieu of fractional shares.

With the split, Chrysler said 13.2 mln shares remain to be purchased in its stock repurchase program that began in late 1984. That program now has a target of 56.3 mln shares with the latest stock split.

Chrysler said in a statement the actions "reflect not only our outstanding performance over the past few years but also our optimism about the company's future."

Min, Zhang, Wright, Ma, CIKM 2010.
Other Data/Applications: Lyrics and Music Separation

Songs (STFT) = Low-rank (music) + Sparse (voices)

Po-Sen Huang, Scott Chen, Paris Smaragdis, Mark Hasegawa-Johnson, ICASSP 2012.
Other Data/Applications: Internet Traffic Anomalies

Network Traffic = Normal Traffic + Sparse Anomalies + Noise

\[ D = L + RS + N \]

Other Data/Applications: View-Invariant Gait Recognition

Same gait from different views

Perspective distortion rectified

Kusakunniran, Wu, Zhang, Ma and Li, submitted to Pattern Recognition 2012.
CONCLUSIONS – *A Unified Theory for Sparsity and Low-Rank*

<table>
<thead>
<tr>
<th></th>
<th><strong>Sparse Vector</strong></th>
<th><strong>Low-Rank Matrix</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-dimensionality of</td>
<td>individual signal</td>
<td>correlated signals</td>
</tr>
<tr>
<td>Measure</td>
<td>$L_0$ norm $|x|_0$</td>
<td>$\text{rank}(X)$</td>
</tr>
<tr>
<td>Convex Surrogate</td>
<td>$L_1$ norm $|x|_1$</td>
<td>Nuclear norm $|X|_*$</td>
</tr>
<tr>
<td>Compressed Sensing</td>
<td>$y = Ax$</td>
<td>$Y = A(X)$</td>
</tr>
<tr>
<td>Error Correction</td>
<td>$y = Ax + e$</td>
<td>$Y = A(X) + E$</td>
</tr>
<tr>
<td>Domain Transform</td>
<td>$y \circ \tau = Ax + e$</td>
<td>$Y \circ \tau = A(X) + E$</td>
</tr>
<tr>
<td>Mixed Structures</td>
<td>$Y = A(X) + B(E) + Z$</td>
<td></td>
</tr>
</tbody>
</table>

Joint NSF Project with Candes and Wright, 2010 - 2014
A Unified THEORY – A Suite of Powerful Regularizers

For compressive robust recovery of a family of low-dimensional structures:

- [Zhou et. al. ’09] Spatially contiguous sparse errors via MRF
- [Bach ’10] – 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

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

- [Wright+Ganesh+Min+Ma, ISIT’12] – superposition of decomposable structures

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

Take home message: Let the data and application tell you the structure...
A Perfect Storm in the Cloud...

Mathematical Theory
(high-dimensional statistics, convex geometry
measure concentration, combinatorics…)

Massive Data
(images, videos,
texts, audios,
speeches, stocks,
user rankings…)

Cloud Computing
(parallel, distributed, networked)

Applications & Services
(data processing,
analysis, compression,
knowledge discovery,
search, recognition…)

Computational Methods
(convex optimization, first-order algorithms,
random sampling, approximate solutions…)

(a) Robust PCA, Random Signs
REFERENCES + ACKNOWLEDGEMENT

Core References:

• **Robust Principal Component Analysis?** Candes, Li, Ma, Wright, Journal of the ACM, 2011.
• **TILT: Transform Invariant Low-rank Textures**, Zhang, Liang, Ganesh, and Ma, IJCV 2012.
• **Compressive Principal Component Pursuit**, Wright, Ganesh, Min, and Ma, ISIT 2012.

More references, codes, and applications on the website:

[http://perception.csl.illinois.edu/matrix-rank/home.html](http://perception.csl.illinois.edu/matrix-rank/home.html)

Colleagues:

• Prof. Emmanuel Candes (Stanford)
• Prof. John Wright (Columbia)
• Prof. Zhouchen Lin (Peking University)
• Dr. Yasuyuki Matsushita (MSRA)
• Dr. Arvind Ganesh (IBM Research, India)
• Prof. Shuicheng Yan (Na. Univ. Singapore)
• Prof. Jian Zhang (Sydney Tech. Univ.)
• Prof. Lei Zhang (HK Polytech Univ.)

Students:

• Zhengdong Zhang (MSRA, Tsinghua University)
• Xiao Liang (MSRA, Tsinghua University)
• Xin Zhang (MSRA, Tsinghua University)
• Kerui Min (UIUC)
• Zhihan Zhou (UIUC)
• Hossein Mobahi (UIUC)
• Guangcan Liu (UIUC)
• Xiaodong Li (Stanford)