Some Recent Advances in the Theory of Low-rank Modeling

Emmanuel Candès

Machine Learning Summer School (MLSS 2001)
Carcans Maubuisson, September 2011
Objective

- Explosion of research on theory of low-rank modeling
- Our goal is to discuss some recent works
  - Some of it is ours
  - Some of it is not
Agenda

1. Matrix completion
2. Robust principal component analysis
Matrix Completion
The Netflix problem

- Netflix database
  - About half a million users
  - About 18,000 movies
- People rate movies
- Sparsely sampled entries
The Netflix problem

- Netflix database
  - About half a million users
  - About 18,000 movies
- People rate movies
- Sparsely sampled entries

![Matrix diagram]
The Netflix problem

- Netflix database
  - About half a million users
  - About 18,000 movies
- People rate movies
- Sparsely sampled entries

Challenge
Complete the “Netflix matrix”

Many such problems → collaborative filtering, partially filled out surveys...
Global positioning from local distances

- Points \( \{x_j\}_{1 \leq j \leq n} \in \mathbb{R}^d \)
- Partial information about distances
  \[ L_{ij} = \| x_i - x_j \|^2 \]

Example (Singer, Biswas et al.)
- Low-powered wirelessly networked sensors
- Each sensor can construct a distance estimate from nearest neighbor
Global positioning from local distances

- Points \( \{x_j\}_{1 \leq j \leq n} \in \mathbb{R}^d \)
- Partial information about distances

\[
L_{ij} = \|x_i - x_j\|^2
\]

Example (Singer, Biswas et al.)
- Low-powered wirelessly networked sensors
- Each sensor can construct a distance estimate from nearest neighbor

Problem
Locate the sensors
Other problems of this kind

- Linear system identification (Vandenberghe et al.)
- Quantum-state tomography (Gross et al.)
- Partially observed covariance matrix (Vaidyanathan et al.)
- Low-rank matrix completion in machine learning (Srebro et al. Vert et al.)
- Structure-from-motion problem in computer vision (Tomasi et al.)
- ...
Matrix completion

- Matrix $L \in \mathbb{R}^{n_1 \times n_2}$
- Observe subset of entries
- Can we guess the missing entries?
Matrix completion

- Matrix $L \in \mathbb{R}^{n_1 \times n_2}$
- Observe subset of entries
- Can we guess the missing entries?

\[
\begin{pmatrix}
\times & ? & ? & ? & \times & ? \\
? & ? & \times & \times & ? & ? \\
\times & ? & ? & \times & ? & ? \\
? & ? & \times & \times & ? & ? \\
? & ? & \times & \times & ? & ? \\
\end{pmatrix}
\]

*Everybody would agree this looks impossible*
Massive high-dimensional data

Engineering/scientific applications: unknown matrix has often (approx.) low rank
Low-rank matrix completion?

Engineering/scientific applications: unknown matrix has often (approx.) low rank

1. Netflix matrix
2. Sensor-net matrix: \[ \| x_i - x_j \|^2, \{ x_i \} \in \mathbb{R}^d \]
   - rank 2 if \( d = 2 \)
   - rank 3 if \( d = 3 \)
   - ...
3. Many others (e.g. quantum-state tomography, computer vision, system id, ...)

\[
\begin{bmatrix}
\times & \times & \times \\
\times & \times & \times \\
\times & \times & \times \\
\end{bmatrix}
\]
Announcing a Joint Seminar of the Committee on Applied and Theoretical Statistics and the Committee on National Statistics of The National Academies...

The Story of the Netflix Prize

Friday, November 4, 2011 • 3:00–5:00 pm
Reception to Follow

Keck Center of the National Academies, Room 100
500 Fifth Street NW
Washington, DC 20001

Just over five years ago, Netflix released more than 100 million movie ratings as part of a data analysis contest to improve methods for recommending movies to customers based on ratings they had provided for previously rented movies. A prize of $1 million was offered for a “recommender” algorithm that outperformed the existing Netflix system Cinematch° by at least 10% in terms of root mean squared prediction error. In a textbook example of “crowdsourcing,” more than 20,000 teams from over 150 countries submitted algorithms. By August 2009, after almost three years of effort, two teams, BellKor’s Pragmatic Chaos and The Ensembl, had surpassed the 10% goal in a finish worthy of its own movie.

Bob Bell (BellKor’s Pragmatic Chaos) and Lester Mackey (The Ensembl) will describe the overall set-up of the competition, the challenges it posed, the main ideas underlying their recommender algorithms, and the interaction among the leading competitors. Emmanuel Candès will then discuss the research avenues stimulated by the various algorithms developed in this competition, some of the resulting advances, and some difficult problems that remain.

— Open to the Public • Please RSVP! —
For planning and building check-in purposes, please RSVP by October 31 to Agnes Gaskin at agaskin@nas.edu or (202) 334-3096.
Low-rank matrix completion?

$L$: $n_1 \times n_2$ matrix of rank $r$

- Singular value decomposition: $L = U \Sigma V^*$

$L$ depends upon $(n_1 + n_2 - r)r$ degrees of freedom < ambient dimension
Low-rank matrix completion?

$L$: $n_1 \times n_2$ matrix of rank $r$

- Singular value decomposition: $L = U \Sigma V^*$

- $L$ depends upon $(n_1 + n_2 - r)r$ degrees of freedom $< \text{ambient dimension}$

Do we need to see all the entries to recover $L$?
Which entries do we get to see?

Rank-1 matrix $L = xy^*$

$$L_{ij} = x_i y_j$$

If single row (or column) is not sampled $\rightarrow$ recovery is not possible
Which entries do we get to see?

Rank-1 matrix \( L = xy^* \)

\[
L_{ij} = x_i y_j
\]

If single row (or column) is not sampled \( \rightarrow \) recovery is not possible

What happens for almost all sampling sets?

\( m \) entries selected uniformly at random \( \rightarrow \Omega_{\text{obs}} \)
Which matrices can we complete?

\[ L = e_1 e^*_n = \begin{bmatrix}
0 & 0 & \cdots & 0 & 1 \\
0 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 0 \\
\end{bmatrix} \]

Cannot be recovered from a small set of entries
Which matrices can we complete?

\[ L = \begin{bmatrix}
* & * & 0 & \cdots & 0 & 0 \\
* & * & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 \\
\end{bmatrix} \]

Cannot be recovered from a small set of entries
Which matrices can we complete?

\[ L = e_1 x^* = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_{n-1} & x_n \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \]

Cannot be recovered from a small set of entries
Which matrices can we complete?

\[ L = e_1 x^* = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_{n-1} & x_n \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \]

Cannot be recovered from a small set of entries

*Intuition: column and row spaces cannot be aligned with basis vectors*
Coherence

\[ L \in \mathbb{R}^{n \times n} = U \Sigma V^* \quad r = \text{rank}(L) \]

Coherence parameter \( \mu \geq 1 \) (C. and Recht '08): for all \( e_i = (0, \ldots, 0, 1, 0 \ldots, 0) \)

\[ \|U^* e_i\|^2 \leq \frac{\mu r}{n} \quad \|V^* e_i\|^2 \leq \frac{\mu r}{n} \]

and

\[ |UV^*|^2_{ij} \leq \frac{\mu r}{n^2} \]

Roughly: small value of \( \mu \rightarrow \) sing. vectors not sparse

Condition holds if \( |U_{ij}|^2 \vee |V_{ij}|^2 \leq \mu/n \)

Random plane of dimension \( r \geq \log n \)

\[ \max_i \|U^* e_i\|^2 \leq O(1)r/n \]
What is information theoretically possible?

C. and Tao (09)

Roughly, no method whatsoever can succeed

\[ m \lesssim \mu \times nr \times \log n \approx df \times \mu \log n \]

For rectangular matrices \( n = \max \dim \)

- Fundamental role played by coherence parameter
- With \( \mu = O(1) \) (incoherence), need \( m \gtrsim nr \log n \)
Recovery algorithm

Hope: only one low-rank matrix consistent with the sampled entries

Recovery by minimum complexity

\[
\begin{align*}
\text{minimize} & \quad \text{rank}(\hat{L}) \\
\text{subject to} & \quad \hat{L}_{ij} = L_{ij} \quad (i, j) \in \Omega_{\text{obs}}
\end{align*}
\]

NP-hard: not feasible for \( n > 10 \)!
Recovery algorithm

Hope: only one low-rank matrix consistent with the sampled entries

Recovery by nuclear-norm minimization (SDP)

\[
\begin{align*}
\text{minimize} & \quad \|L\|_* = \sum_{i=1}^{n} \sigma_i(L) \\
\text{subject to} & \quad \hat{L}_{ij} = L_{ij} \quad (i, j) \in \Omega_{\text{obs}}
\end{align*}
\]

- Convex relaxation of the rank minimization program
- Ball \( \{X : \|X\|_* \leq 1\} \): convex hull of rank-1 matrices obeying \( \|xy^*\| \leq 1 \)
Recovery algorithm

Hope: only one low-rank matrix consistent with the sampled entries

Recovery by nuclear-norm minimization (SDP)

\[
\begin{align*}
\text{minimize} & \quad \| \hat{L} \|_* = \sum_{i=1}^{n} \sigma_i(\hat{L}) \\
\text{subject to} & \quad \hat{L}_{ij} = L_{ij} \quad (i, j) \in \Omega_{\text{obs}}
\end{align*}
\]

- Convex relaxation of the rank minimization program
- Ball \( \{ X : \| X \|_* \leq 1 \} \): convex hull of rank-1 matrices obeying \( \| xy^* \| \leq 1 \)

Trace norm heuristics
- Mesbahi & Papavassilopoulos '97
- Beck & D’Andrea '98
- Fazel '02
Near-optimal matrix completion

\[
\begin{align*}
\text{minimize} & \quad \| \hat{L} \|_* \\
\text{subject to} & \quad \hat{L}_{ij} = L_{ij} \quad (i,j) \in \Omega_{\text{obs}}
\end{align*}
\]

Theorem (C. and Tao ’09 improving C. and Recht ’08)

- \( \text{rank}(L) = r \)
- \( \Omega_{\text{obs}} \) random set of size \( m \)

Solution to SDP is exact with probability at least \( 1 - n^{-10} \) if

\[
m \gtrsim \mu nr \log^a n \quad a \leq 6 \quad (\text{sometimes 2})
\]

Gross’ near-optimal improvement

\[
m \gtrsim \mu nr \log^2 n
\]
Related work

- Related results
  - Recht Parrilo Fazel '07
  - Keshavan, Oh and Montanari '09

- Earlier result [C. and Recht '08]:
  \[ m \gtrsim \mu n^{6/5} r \log n \]

- Other contributions
  - Cai, C. and Shen '08
  - Mazumder, Hastie and Tibshirani '09
  - Ma and Goldfarb '09
  - ...
minimize $\|\hat{L}\|_*$
subject to $\hat{L}_{ij} = L_{ij}$ $(i, j) \in \Omega$
Geometry

minimize $\|\hat{L}\|_*$

subject to $\hat{L}_{ij} = L_{ij} \ (i, j) \in \Omega$

Feasible set
General formulation

- $A_1, \ldots, A_N$ (orthonormal) basis of $\mathbb{R}^{n \times n}$ ($N = n^2$)
- $\Omega \subset \{1, \ldots, N\}$

minimize $\|X\|_*$
subject to $\langle A_k, X \rangle = \langle A_k, L \rangle$ $k \in \Omega$

If incoherence between sensing matrices $\{A_k\}$ and col. + row space everything should work...
Two orthonormal bases $F = [f_1, \ldots, f_n]$, $G = [g_1, \ldots, g_n]$

Orthobasis of $n \times n$ matrices: $\{f_i g_j^*\}_{1 \leq i, j \leq n}$

minimize $\|X\|_*$
subject to $f_i^* X g_j = f_i^* L g_j$ $(i, j) \in \Omega$

Succeeds if col. (resp. row) space of $L$ incoherent with $\{f_i\}$ (resp. $\{g_i\}$)
Two orthonormal bases $F = [f_1, \ldots, f_n]$, $G = [g_1, \ldots, g_n]$

Orthobasis of $n \times n$ matrices: $\{f_i g_j^*\}_{1 \leq i, j \leq n}$

minimize $\|X\|_*$

subject to $f_i^* X g_j = f_i^* L g_j$ $(i, j) \in \Omega$

Succeeds if col. (resp. row) space of $L$ incoherent with $\{f_i\}$ (resp. $\{g_i\}$)

Why? Because $f_i^* X g_j = e_i^* (F^* X G) e_j$
Quantum-state tomography

- $k$: spin-1/2 system in an *unknown* quantum state $L \in \mathbb{C}^{n \times n}$ (density matrix)
  
  $$n = 2^k, \quad \text{trace}(L) = 1, \quad L \succeq 0$$

- Quantum measurements (data)
  
  $\mathbb{E} [\text{measurement with observable } A_j] = \langle A_j, L \rangle = \text{trace}(A_j^* L)$

  e.g. $\{A_j\}$: tensor Pauli matrices

**Q?** Can we reduce # measurements by using the structure of special classes of quantum states?

- pure state $\rightarrow \text{rank}(L) = 1$

- interesting mixed states $\rightarrow$ (approx) low rank
Quantum-state tomography

- $k$: spin-$1/2$ system in an unknown quantum state $L \in \mathbb{C}^{n \times n}$ (density matrix)
  \[ n = 2^k, \quad \text{trace}(L) = 1, \quad L \succeq 0 \]

- Quantum measurements (data)
  \[ \mathbb{E}[\text{measurement with observable } A_j] = \langle A_j, L \rangle = \text{trace}(A_j^*L) \]
  e.g. \{A_j\}: tensor Pauli matrices

Q? Can we reduce \# measurements by using the structure of special classes of quantum states?
- pure state $\rightarrow \text{rank}(L) = 1$
- interesting mixed states $\rightarrow$ (approx) low rank

A. Yes. Sample in proportion to the rank of the quantum state (Gross '09)
General statement

$A_1, \ldots, A_{n^2}$ (orthonormal) basis of $\mathbb{R}^{n \times n}$ and observe ($L = U \Sigma V^*$)

$$y_k = \langle A_k, L \rangle \quad k \in \Omega$$

$\Omega$ random set of size $m$

- Coherence assumption

$$\max_k \| P_U A_k \|_F^2 \leq \mu r / n \quad \max_k \| A_k P_V \|_F^2 \leq \mu r / n$$

- At least one of the two conditions

$$\max_k \| A_k \|_F^2 \leq \mu / n$$

$$\max_k |\langle A_k, UV^* \rangle|^2 \leq \mu r / n^2$$

Theorem (Gross '09)

Min nuclear-norm solution is exact with high prob. provided

$$m \gtrsim \mu \times nr \times \log_2 n$$
General statement

$A_1, \ldots, A_{n^2}$ (orthonormal) basis of $\mathbb{R}^{n \times n}$ and observe ($L = U \Sigma V^*$)

$$y_k = \langle A_k, L \rangle \quad k \in \Omega$$

$\Omega$ random set of size $m$

- Coherence assumption

$$\max_k \|P_U A_k\|_F^2 \leq \mu r / n \quad \max_k \|A_k P_V\|_F^2 \leq \mu r / n$$

- At least one of the two conditions

$$\max_k \|A_k\|^2 \leq \mu / n$$

$$\max_k |\langle A_k, UV^* \rangle|^2 \leq \mu r / n^2$$

Theorem (Gross ’09)

Min nuclear-norm solution is exact with high prob. provided

$$m \geq \mu \times nr \times \log^2 n$$
Robust PCA
Matrix completion from noisy entries

\[ Y_{ij} = L_{ij} + Z_{ij}, \quad (i, j) \in \Omega_{\text{obs}} \quad Z_{ij} \text{ iid } \mathcal{N}(0, \sigma^2) \]

Recovery by SDP with relaxed constraints

\[
\begin{align*}
\text{minimize} & \quad \|\hat{L}\|_* \\
\text{subject to} & \quad \sum_{i,j \in \Omega_{\text{obs}}} (\hat{L}_{ij} - Y_{ij})^2 \leq (1 + \epsilon)n\sigma^2
\end{align*}
\]
Matrix completion from noisy entries

\[ Y_{ij} = L_{ij} + Z_{ij}, \quad (i, j) \in \Omega_{\text{obs}} \quad Z_{ij} \text{ iid } \mathcal{N}(0, \sigma^2) \]

Recovery by SDP with relaxed constraints

\[
\begin{align*}
\text{minimize} & \quad \| \hat{L} \|_* \\
\text{subject to} & \quad \sum_{i,j \in \Omega_{\text{obs}}} (\hat{L}_{ij} - Y_{ij})^2 \leq (1 + \epsilon)n\sigma^2 \\
\end{align*}
\]

Theorem (C. and Plan, ’09)

Same assumptions as before. With very high prob.

\[ n^{-2}\| \hat{L} - L \|_F^2 \lesssim n\sigma^2 \]

When exact recovery occurs, noisy variant is stable
Matrix completion from noisy entries

\[ Y_{ij} = L_{ij} + Z_{ij}, \quad (i, j) \in \Omega_{\text{obs}} \quad Z_{ij} \text{ iid } \mathcal{N}(0, \sigma^2) \]

Recovery by SDP with relaxed constraints

\[
\begin{align*}
\text{minimize} & \quad \|\hat{L}\|_* \\
\text{subject to} & \quad \sum_{i,j \in \Omega_{\text{obs}}} (\hat{L}_{ij} - Y_{ij})^2 \leq (1 + \epsilon)n\sigma^2
\end{align*}
\]

Theorem (C. and Plan, ’09)

Same assumptions as before. With very high prob.

\[ n^{-2}\|\hat{L} - L\|_F^2 \lesssim n\sigma^2 \]

When exact recovery occurs, noisy variant is stable

Some other works

- Koltchinskii, Lounici & Tsybakov (’10)
- Bunea, She & Wegkamp (’10)
- Negahban & Wainwright (’10)
- Rohde & Tsybakov (’10)
Gross errors

Observe corrupted samples from $L + E$

- $L$ low-rank matrix
- $E$ entries that have been tampered with – impulsive noise

Goal

Recover $L$: make approach robust vis a vis gross errors
The separation problem

\[ M = L + E \]

- \( M \): data matrix (observed)
- \( L \): low-rank (unobserved)
- \( E \): sparse (unobserved)
The separation problem

\[ M = L + E \]

- \( M \): data matrix (observed)
- \( L \): low-rank (unobserved)
- \( E \): sparse (unobserved)

Problem: can we recover \( L \) and \( E \) accurately?

Again, seems impossible
Classical PCA

\[ M = L + N \]

- \( L \): low-rank (unobserved)
- \( N \): (small) perturbation

Dimensionality reduction (Schmidt 1907, Hotelling 1933)

\[
\begin{align*}
\text{minimize} & \quad \| M - \hat{L} \| \\
\text{subject to} & \quad \text{rank}(\hat{L}) \leq k
\end{align*}
\]

Solution given by truncated SVD

\[ M = U \Sigma V^* = \sum_i \sigma_i u_i v_i^* \quad \Rightarrow \quad \hat{L} = \sum_{i \leq k} \sigma_i u_i v_i^* \]
Classical PCA

\[ M = L + N \]

- \( L \): low-rank (unobserved)
- \( N \): (small) perturbation

Dimensionality reduction (Schmidt 1907, Hotelling 1933)

\[
\begin{align*}
\text{minimize} & \quad \| M - \hat{L} \| \\
\text{subject to} & \quad \text{rank}(\hat{L}) \leq k
\end{align*}
\]

Solution given by truncated SVD

\[
M = U \Sigma V^* = \sum_i \sigma_i u_i v_i^* \implies \hat{L} = \sum_{i \leq k} \sigma_i u_i v_i^*
\]

Fundamental statistical tool: enormous impact
PCA and corruptions/outliers

PCA: very sensitive to outliers
PCA and corruptions/outliers

PCA: very sensitive to outliers

Breaks down with one (badly) corrupted data point
Robust PCA

Gross errors frequently occur in many applications

- Image processing
- Web data analysis
- Bioinformatics
- ...
Example: Face recognition under varying illuminations

Images of same face under varying illuminations $\sim 9D$ harmonic plane (Basri and Jacobs, 03)
Real data are corrupted, have missing blocks $\rightarrow$ classical methods break down

How do we develop provably correct and efficient algorithms for recovery of low-dimensional linear structure from non-ideal observations?
When does separation make sense?

What if $M = L + E$ is both low-rank and sparse?

$$M = e_1 e_n^* = \begin{bmatrix}
0 & 0 & \cdots & 0 & 1 \\
0 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 0
\end{bmatrix}$$
When does separation make sense?

What if $M = L + E$ is both low-rank and sparse?

$$M = e_1 e^*_n = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

Low-rank component cannot be sparse

Will assume $L \in \mathbb{R}^{n \times n}$ obeys previous incoherence condition

“sing. vectors are not sparse”
What if the sparse component has low-rank?

E.g. first column of $E$ is minus that of $L$

$$E = \begin{bmatrix} * & 0 & \cdots & 0 & 0 \\ * & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ * & 0 & \cdots & 0 & 0 \end{bmatrix} \quad \Rightarrow \quad M = L + E = \begin{bmatrix} 0 & * & \cdots & * & * \\ 0 & * & \cdots & * & * \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & * & \cdots & * & * \end{bmatrix}$$
What if the sparse component has low-rank?

E.g. first column of $E$ is minus that of $L$

\[
E = \begin{bmatrix}
* & 0 & \cdots & 0 & 0 \\
* & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
* & 0 & \cdots & 0 & 0
\end{bmatrix}
\Rightarrow M = L + E = \begin{bmatrix}
0 & * & \cdots & * & * \\
0 & * & \cdots & * & * \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & * & \cdots & * & *
\end{bmatrix}
\]

Sparsity pattern will be assumed (uniform) random
Principal Component Pursuit (PCP)

\[ M = L + E \]

- \( L \) unknown (rank unknown)
- \( E \) unknown (\# of entries \( \neq 0 \), locations, magnitudes all unknown)

See also Chandrasekaran, Sanghavi, Parrilo, Willsky ('09)

- **nuclear norm:**
  \[ \|L\|_* = \sum \sigma_i(L) \] (sum of sing. values)

- **\( \ell_1 \) norm:**
  \[ \|S\|_1 = \sum |S_{ij}| \] (sum of abs. values)
**Principal Component Pursuit (PCP)**

\[
M = L + E
\]

- \( L \) unknown (rank unknown)
- \( E \) unknown (\# of entries \( \neq 0 \), locations, magnitudes all unknown)

**Recovery via (convex) PCP**

\[
\begin{align*}
\text{minimize} & \quad \|\hat{L}\|_* + \lambda \|\hat{E}\|_1 \\
\text{subject to} & \quad \hat{L} + \hat{E} = M
\end{align*}
\]

See also Chandrasekaran, Sanghavi, Parrilo, Willsky (’09)

- nuclear norm: \( \|L\|_* = \sum_i \sigma_i(L) \) (sum of sing. values)
- \( \ell_1 \) norm: \( \|S\|_1 = \sum_{ij} |S_{ij}| \) (sum of abs. values)
Main result: $M = L + E$

**Theorem (C., Li, Ma and Wright, 09)**

- $L$ is $n \times n$ of rank $(L) \leq \rho_r n \mu^{-1} (\log n)^{-2}$
- $E$ is $n \times n$, random sparsity pattern of cardinality $m \leq \rho_s n^2$

Then with probability $1 - O(n^{-10})$, PCP with $\lambda = 1/\sqrt{n}$ is exact:

$$\hat{L} = L, \quad \hat{E} = E$$

Same conclusion for rectangular matrices with $\lambda = 1/\sqrt{\text{max dim}}$

- Exact
  - whatever the magnitudes of $L$!
  - whatever the magnitudes of $E$!
- No tuning parameter!
Connections with matrix completion (MC)

Missing vs. corrupted data

\[
\begin{bmatrix}
\times & ? & ? & ? & \times & ? \\
? & ? & \times & \times & ? & ? \\
\times & ? & ? & \times & ? & ? \\
? & ? & \times & ? & ? & \times \\
? & ? & \times & \times & ? & ? \\
\end{bmatrix}
\]

MC: missing

\[
\begin{bmatrix}
\times & \text{skull} & \text{skull} & \text{skull} & \times & \times \\
\text{skull} & \text{skull} & \text{skull} & \text{skull} & \times & \times \\
\times & \text{skull} & \text{skull} & \text{skull} & \times & \times \\
\text{skull} & \text{skull} & \text{skull} & \text{skull} & \times & \times \\
\times & \text{skull} & \text{skull} & \text{skull} & \times & \times \\
\end{bmatrix}
\]

RPCA: corrupted

*Harder to detect and correct than to fill in*
Phase transitions in probability of success

\[ L = XY^* \] is a product of independent \( n \times r \) i.i.d. \( \mathcal{N}(0, 1/n) \) matrices
Other works

Chandrasekaran, Sanghavi, Parrilo and Willsky (09): deterministic results

- Hsu, Kakade and Zhang (10)
- Chen, Jalali, Sanghavi and Caramanis (11)
- Li (11)
**Tying it together**

Theorem (C., Li, Ma and Wright, 09)

$L$ as before, rank $(L) \leq \rho_0 n^{\mu - 1 - 2/\Omega_{\text{obs}}}$. Each observed entry corrupted with prob. $\tau \leq \tau_0$. Then with prob. $1 - O(1/n - 10)$, PCP with $\lambda = 1/\sqrt{n}$ is exact: $\hat{L} = L$. Same conclusion for rectangular matrices with $\lambda = 1/\sqrt{\max \text{dim}}$.

**PCP**

$$\begin{align*}
\text{min} & \quad \|\hat{L}\|_* + \lambda \|\hat{E}\|_1 \\
\text{s. t.} & \quad \hat{L}_{ij} + \hat{E}_{ij} = L_{ij} + E_{ij} \quad (i, j) \in \Omega_{\text{obs}}
\end{align*}$$
Tying it together

**PCP**

\[
\begin{align*}
\min & \quad \|\hat{L}\|_* + \lambda \|\hat{E}\|_1 \\
\text{s. t.} & \quad \hat{L}_{ij} + \hat{E}_{ij} = L_{ij} + E_{ij} \quad (i, j) \in \Omega_{\text{obs}}
\end{align*}
\]

**Theorem (C., Li, Ma and Wright, 09)**

- \( L \) as before, \( \text{rank}(L) \leq \rho_0 n \mu^{-1}(\log n)^{-2} \)
- \( \Omega_{\text{obs}} \) random set of size \( m = 0.1n^2 \) (missing frac. is arbitrary)
- Each observed entry corrupted with prob. \( \tau \leq \tau_0 \)

Then with prob. \( 1 - O(n^{-10}) \), PCP with \( \lambda = 1/\sqrt{0.1n} \) is exact:

\[ \hat{L} = L \]

Same conclusion for rectangular matrices with \( \lambda = 1/\sqrt{0.1\max \text{dim}} \)
Gross errors + noise

Extension (C., Li, Ma, Wright & Zhou ’10)

\[ Y_{ij} = L_{ij} + E_{ij} + Z_{ij} \quad (i, j) \in \Omega \]

- \( L \) low rank
- \( E \) sparse (gross errors)
- \( Z \) stochastic or deterministic perturbation
Gross errors + noise

Extension (C., Li, Ma, Wright & Zhou '10)

\[ Y_{ij} = L_{ij} + E_{ij} + Z_{ij} \quad (i, j) \in \Omega \]

- \( L \) low rank
- \( E \) sparse (gross errors)
- \( Z \) stochastic or deterministic perturbation

PCP with relaxed constraints \( \Rightarrow \) error as if no impulsive noise
Other models: Xu, Caramanis & Sanghavi ’10

Observe all entries of $Y$

\[
Y = L + C (+Z)
\]

- $L$ low rank
- $C$ column of outliers
- $Z$ stochastic or deterministic perturbation

Goal

Achieve segmentation (noiseless case):
- Identify columns in low-dim subspace
- Identify outliers
Computational issues

Wish to solve the SDP

\[
\begin{align*}
\text{minimize} & \quad \|L\|_* + \lambda\|E\|_1 \\
\text{subject to} & \quad L + E = M
\end{align*}
\]

- Off-the-shelf algorithms (SDPT3, SeDuMi) need \( n < 80, 100 \)
- Customized IPMs don’t do much better

Have developed a simple and scalable algorithm via the Alternating Direction Method of Multipliers (ADMM)
## Empirical performance

\[
\begin{array}{|c|c|c|c|c|c|c|c|}
\hline
n & \text{rank}(L) & \|E\|_0 & \text{rank}(\hat{L}) & \|\hat{E}\|_0 & \frac{\|\hat{L} - L\|_F}{\|L\|_F} & \# \text{ SVD} & \text{Time}(s) \\
\hline
500 & 25 & 12,500 & 25 & 12,500 & 1.1 \times 10^{-6} & 16 & 2.9 \\
1,000 & 50 & 50,000 & 50 & 50,000 & 1.2 \times 10^{-6} & 16 & 12.4 \\
2,000 & 100 & 200,000 & 100 & 200,000 & 1.2 \times 10^{-6} & 16 & 61.8 \\
3,000 & 250 & 450,000 & 250 & 450,000 & 2.3 \times 10^{-6} & 15 & 185.2 \\
\hline
\end{array}
\]

\[\text{rank}(L) = 0.05 \times n, \quad \|E\|_0 = 0.05 \times n^2.\]

\[
\begin{array}{|c|c|c|c|c|c|c|c|}
\hline
n & \text{rank}(L) & \|E\|_0 & \text{rank}(\hat{L}) & \|\hat{E}\|_0 & \frac{\|\hat{L} - L\|_F}{\|L\|_F} & \# \text{ SVD} & \text{Time}(s) \\
\hline
500 & 25 & 25,000 & 25 & 25,000 & 1.2 \times 10^{-6} & 17 & 4.0 \\
1,000 & 50 & 100,000 & 50 & 100,000 & 2.4 \times 10^{-6} & 16 & 13.7 \\
2,000 & 100 & 400,000 & 100 & 400,000 & 2.4 \times 10^{-6} & 16 & 64.5 \\
3,000 & 150 & 900,000 & 150 & 900,000 & 2.5 \times 10^{-6} & 16 & 191.0 \\
\hline
\end{array}
\]

\[\text{rank}(L) = 0.05 \times n, \quad \|E\|_0 = 0.10 \times n^2.\]

---

Computational cost higher than classical PCA but not by a large factor!
Implementation status

\( n \times n \) matrices

- Implementation on desktop for \( n \sim 10^3, 10^4 \)

- Distributed implementation for \( n \sim 10^6 \) on Redmond HPC clusters (MSRA)

- Support applications with real high-dim. data
  - images
  - videos
  - audio
  - text documents
  - ...


Some Applications
Application to video surveillance

Sequence of 200 video frames (144 × 172 pixels) with a static background

Problem: detect any activity in the foreground

RPCA
Background modeling from surveillance video, I

Alternating minimization of an M-estimator (De La Torre and Black, '03)
Background modeling from surveillance video, II

Three frames from a 250 frame sequence taken in a lobby, with varying illumination (Li et al., '04).
APPLICATIONS – Repairing vintage movies

Original $D$  

Repaired $A$  

Corruptions

Frame 1

480×620 pixels
APPLICATIONS – Repairing vintage movies

Original \( D \)  

Repaired \( A \)

Corruptions

Frame 2
APPLICATIONS – Repairing vintage movies

Original $D$  Repaired $A$

Corruptions

Frame 3
APPLICATIONS – Repairing vintage movies

Original $D$  Repaired $A$

Corruptions

Frame 4
APPLICATIONS – Repairing vintage movies

Original $D$  

Repaired $A$

Corruptions

Frame 5
APPLICATIONS – Repairing vintage movies

Original $D$

Repaired $A$

Corruptions

Frame 6
APPLICATIONS – Repairing vintage movies

Original $D$  Repaired $A$

Corruptions

Frame 7
APPLICATIONS – Faces under varying illumination

58 images of one person under varying lighting:

[Candes, Li, Ma, and Wright, Journal of the ACM, May 2011.]
APPLICATIONS – *Faces under varying illumination*

58 images of one person under varying lighting:

*Candes, Li, Ma, and Wright, Journal of the ACM, May 2011.*
Robust batch image alignment (Ma et al.)

- **Input**: $M$ corrupted and misaligned batch of images (data)
- **Output**: $L$ aligned low-rank images; $S$ sparse errors

(Model) \[ M \circ \tau = L_0 + S_0 \]

$\tau$: parametric deformation (rigid, affine, projective)
Robust batch image alignment (Ma et al.)

- **Input:** $M$ corrupted and misaligned batch of images (data)
- **Output:** $L$ aligned low-rank images; $S$ sparse errors

(Model) \[ M \circ \tau = L_0 + S_0 \]

$\tau$: parametric deformation (rigid, affine, projective)

Bootstrap: find $L$ and $S$ and $\tau$ solution to

\[
\begin{align*}
\text{minimize} & & \|L\|_* + \lambda \|S\|_1 \\
\text{subject to} & & L + S = M \circ \tau
\end{align*}
\]
APPLICATIONS – 2D image matching and 3D modeling

\[ D \circ \tau \]

\[ \tau \in 2D \text{ homographies} \]

Peng, Ganesh, Wright, Ma, CVPR’10
### APPLICATIONS – *Batch face alignment: accuracy evaluation*

<table>
<thead>
<tr>
<th></th>
<th>Mean error</th>
<th>Error std.</th>
<th>Max error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial misalignment</td>
<td>2.5</td>
<td>1.03</td>
<td>4.87</td>
</tr>
<tr>
<td>Vedaldi (direct/gradient)</td>
<td>1.97/1.66</td>
<td>1.11/0.85</td>
<td>5.71/4.02</td>
</tr>
<tr>
<td>R A S L (this work)</td>
<td>0.48</td>
<td>0.23</td>
<td>1.07</td>
</tr>
</tbody>
</table>

100 misaligned corrupted images:

Vedaldi CVPR’08 direct/gradient

R A S L:
APPLICATIONS – Simultaneous Alignment and Repairing

\[ D \circ \tau \]

\[ A \]

\[ E \]

Peng, Ganesh, Wright, Ma, CVPR’10
APPLICATIONS – Celebrities from the Internet

Average face before alignment & repairing

Gloria Macapagal Arroyo
Jennifer Capriati
Laura Bush
Serena Williams
Barack Obama
Ariel Sharon
Arnold Schwarzenegger
Colin Powell
Donald Rumsfeld
George W Bush
Gerhard Schroeder
Hugo Chavez
Jacques Chirac
Jean Chretien
John Ashcroft
Junichiro Koizumi
Lleyton Hewitt
Luiz Inacio Lula da Silva
Tony Blair
Vladimir Putin

Peng, Ganesh, Wright, Ma, CVPR’10
APPLICATIONS – *Face recognition with less controlled data?*
APPLICATIONS – Aligning handwritten digits

$D$

Learned-Miller PAMI’06

Vedaldi CVPR’08

$D \circ \tau$

$A$

$E$
The world we see (through camera) is tilted!

augmented reality

world lens

3D map
Problem: Given $D \circ \tau = A_0 + E_0$, recover $\tau, A_0, E_0$

Solution: estimate the deformation and low-rank texture simultaneously

Iteratively solving the linearized convex program:

$$\min \| A \|_* + \lambda \| E \|_1 \quad \text{subj} \quad A + E = D \circ \tau_k + J\Delta \tau$$
TILT via Iterative RPCA-Like Convex Optimization

Iteration Processes

\[ D \quad D \circ \tau \quad A \quad E \]
TILT: Examples of Symmetric Patterns and Textures

Input (red window)

Output (rectified green window)
TILT – Robust to Background, Occlusion, and Corruption

$D$  $D \circ \tau$  $A$  $E$

Zhang, Liang, Ganesh, Ma, ACCV’10
TILT: All Types of Regular Geometric Structures in Images

- an ideal edge
- an ideal corner
- symmetry
- man-made

Rectified Low-rank Textures

Zhang, Liang, Ganesh, Ma, ACCV’10
TILT: Examples of Characters, Signs, and Texts

Input (red window)

Output (rectified green window)
TILT: More Examples

Input (red window)

Output (rectified green window)
TILT – 3D Geometry from a Single Image

Zhang, Liang, Ganesh, Ma, ACCV’10
TILT Applications: Augmented Reality
Other 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

Low dimensional topic models with keywords...

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.”

\[ d_{ij} \] word frequency (or TF-IDF)
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."
Summary

- Lots of exciting work in theory of low-rank models (matrix completion)
- Lots more needs to be done