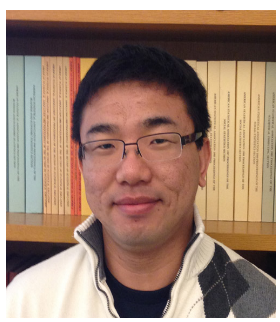
Low-dimensional Structures and Deep Models for High-dimensional (Visual) Data

Yi Ma

EECS Department, UC Berkeley

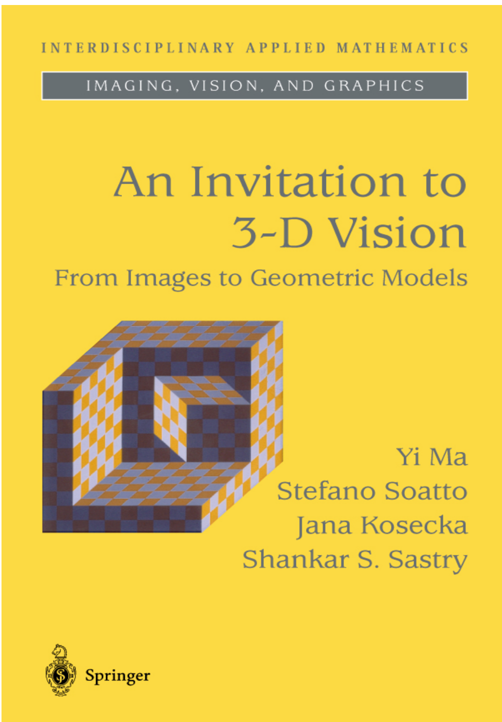
     

Dutch Mathematical Congress, April 4, 2018

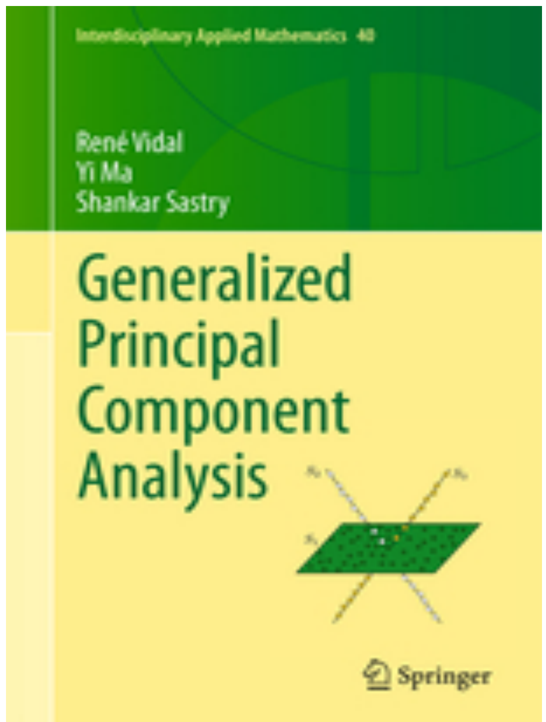
My Interests – From 3-D Vision to High-Dim Data

In order to recover 3D geometry from 2D images, we need to understand low-dim structures in high-dim spaces…

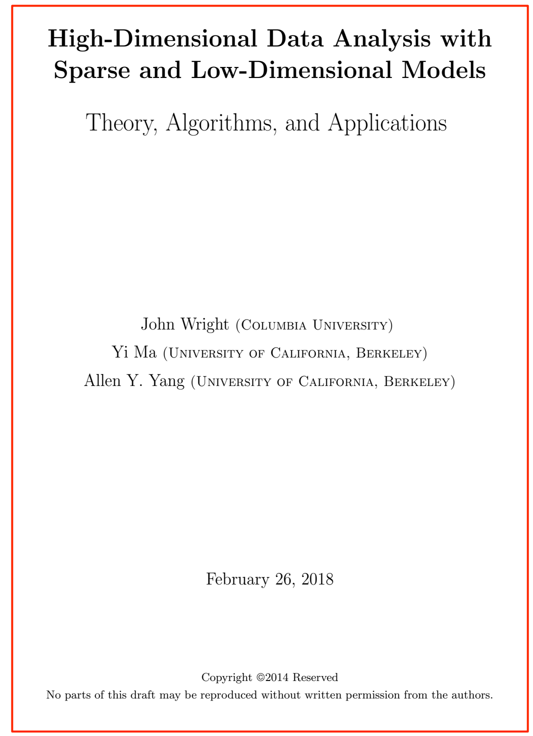
2003



2016



soon


High-Dimensional Data Analysis with Sparse and Low-Dimensional Models

Theory, Algorithms, and Applications

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

February 26, 2018

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

Capturing Shape and Texture of 3D Objects

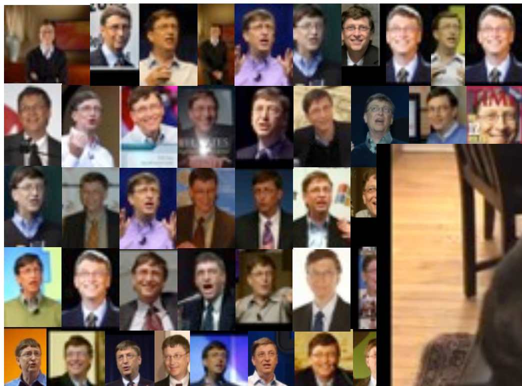
Shanghai Museum Items

On HTC VIVE

On iPhone VR kit

With Jingyi Yu of ShanghaiTech, 2017

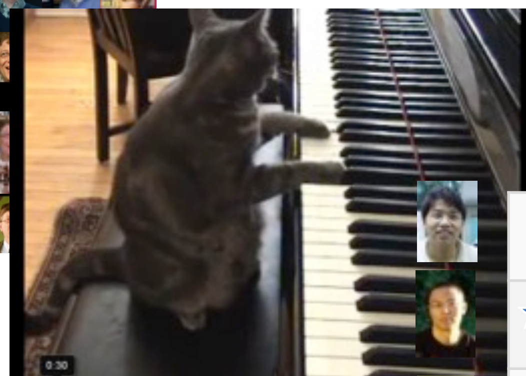
CONTEXT – Data increasingly massive, high-dimensional…



Images

⇓➢ ***1M pixels***

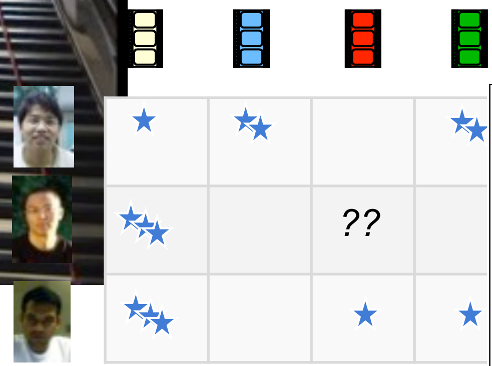
Compression
De-noising
Super-resolution
Recognition…



Videos

⇓➢ ***1B voxels***

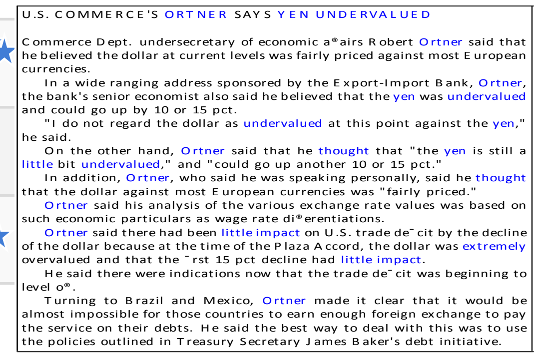
Streaming
Tracking
Stabilization…



User data

⇓➢ ***1B users***

Clustering
Classification
Collaborative filtering…



U.S. COMMERCE'S ORTNER SAYS YEN UNDERVALUED

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

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

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

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

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

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

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

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

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

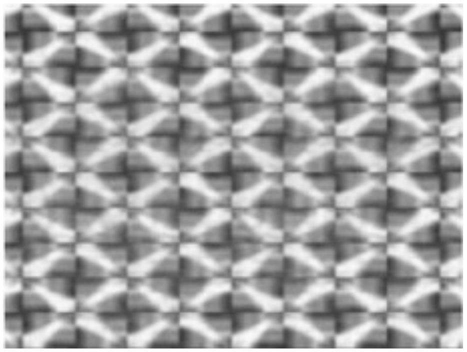
Web data

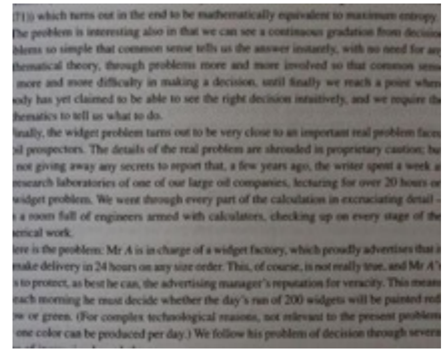
⇓➢ ***100B webpages***

Indexing
Ranking
Search…

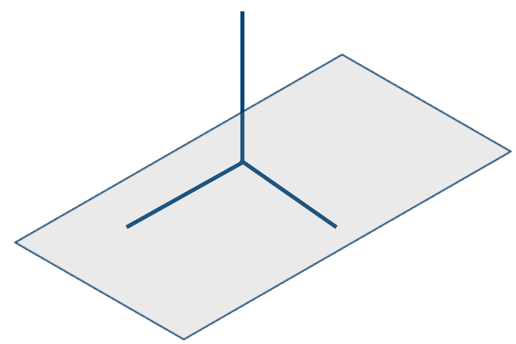
How to extract low-dim structures from such high-dim data?

CONTEXT – Low dimensional structures in visual data




(71)) which turns out in the end to be mathematically equivalent to maximum entropy
The problem is interesting also in that we can see a continuous gradation from decision
blems so simple that common sense tells us the answer instantly, with no need for any
thematical theory, through problems more and more involved so that common sense
more and more difficulty in making a decision, until finally we reach a point where
body has yet claimed to be able to see the right decision intuitively, and we require the
thematics to tell us what to do.
Finally, the widget problem turns out to be very close to an important real problem facing
oil prospectors. The details of the real problem are shrouded in proprietary caution; but
not giving away any secrets to report that, a few years ago, the writer spent a week at
research laboratories of one of our large oil companies, lecturing for over 20 hours on
widget problem. We went through every part of the calculation in excruciating detail –
a room full of engineers armed with calculators, checking up on every stage of the
nerical work.
Here is the problem: Mr A is in charge of a widget factory, which proudly advertises that it
make delivery in 24 hours on any size order. This, of course, is not really true, and Mr A's
is to protect, as best he can, the advertising manager's reputation for veracity. This means
each morning he must decide whether the day's run of 200 widgets will be painted red,
ow or green. (For complex technological reasons, not relevant to the present problem,
one color can be produced per day.) We follow his problem of decision through several

Visual data exhibit ***low-dimensional structures*** due to rich ***local*** regularities, ***global*** symmetries, ***repetitive*** patterns, or ***redundant*** sampling.



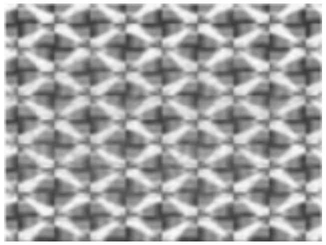
CONTEXT – PCA: Fitting Data with a Low-dim. Subspace

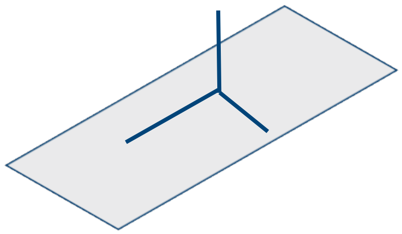
If we view the data (image) as a matrix

$$A = [\mathbf{a}_1 \mid \cdots \mid \mathbf{a}_n] \in \mathbb{R}^{m \times n}$$

then

$$r \doteq \operatorname{rank}(A) \ll m.$$



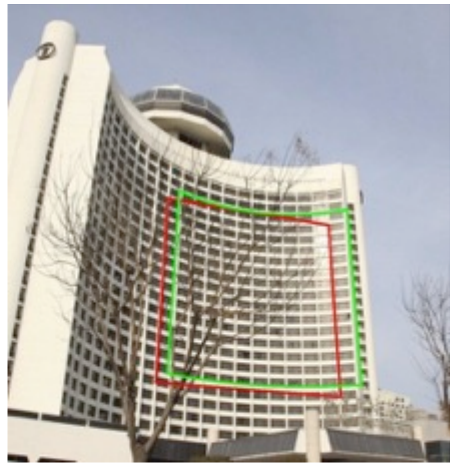


Principal Component Analysis (PCA) via singular value decomposition (SVD):

- 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.**

CONTEXT – But life is not so easy…



*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 probailitatum que respondent diversis valoribus summe errorum post plures observationes … , before 1756*

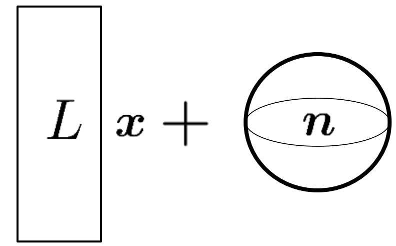
A. Legendre. *Nouvelles methodes pour la determination des orbites des cometes*, 1806

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

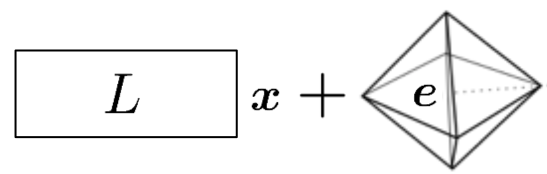
A. Beurling. *Sur les integrales de Fourier absolument convergentes et leur application a une transformation functionelle*, 1938

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

⋮



over-determined
+ dense, Gaussian

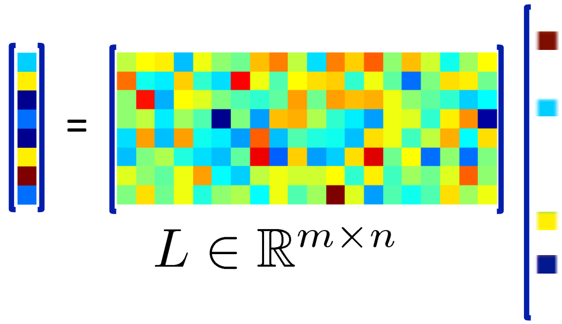


underdetermined
+ sparse, Laplacian

CONTEXT – Recent related progress

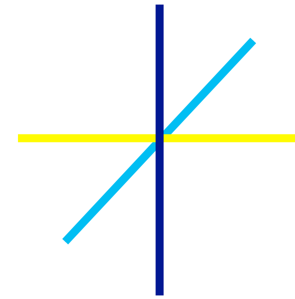
Sparse recovery: Given $y = L{\color{red}x_0}$, $L \in \mathbb{R}^{m \times n}$, $m \ll n$, recover ${\color{red}x_0}$.

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



$L \in \mathbb{R}^{m \times n}$

${\color{red}x} \in \mathbb{R}^n$



Impossible in general ($m \ll n$)

Well-posed if ${\color{red}x_0}$ is structured *(sparse),* but still **NP-hard**

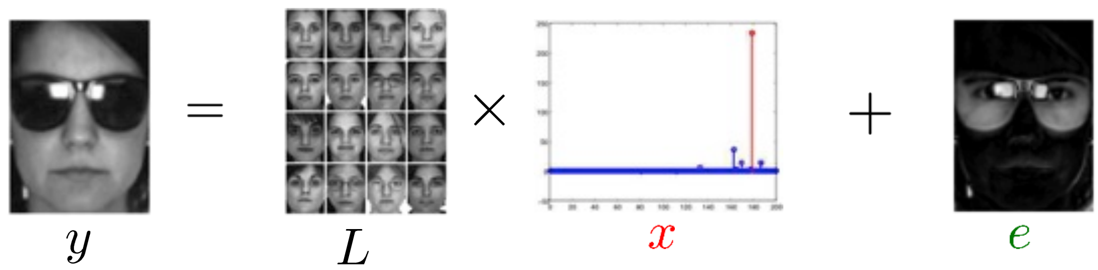
Tractable via convex optimization: $\min \|{\color{red}x}\|_1 \text{ s.t. } y = L{\color{red}x}$

… 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, Dimakis+Xu+Hassibi'09, … and many others*

CONTEXT – Recent related progress

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)*

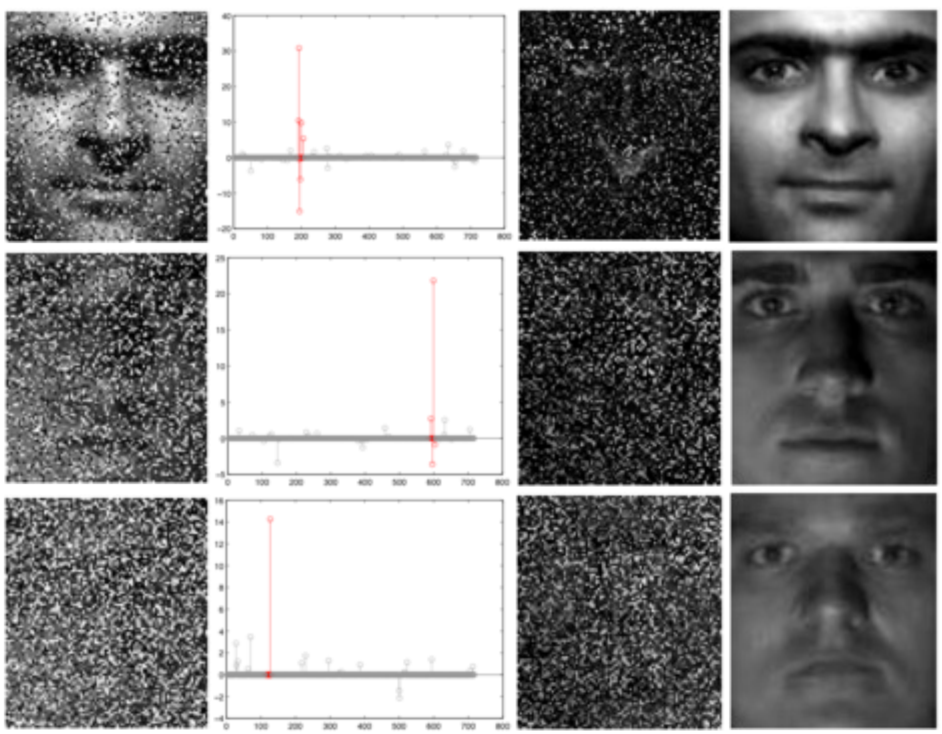
Hugely active area*: Candès+Tao'05, Wright+Ma'10, Nguyen+Tran'11, Li '11, also Zhang, Yang, Huang'11, Oymak+Tropp'15 etc…*

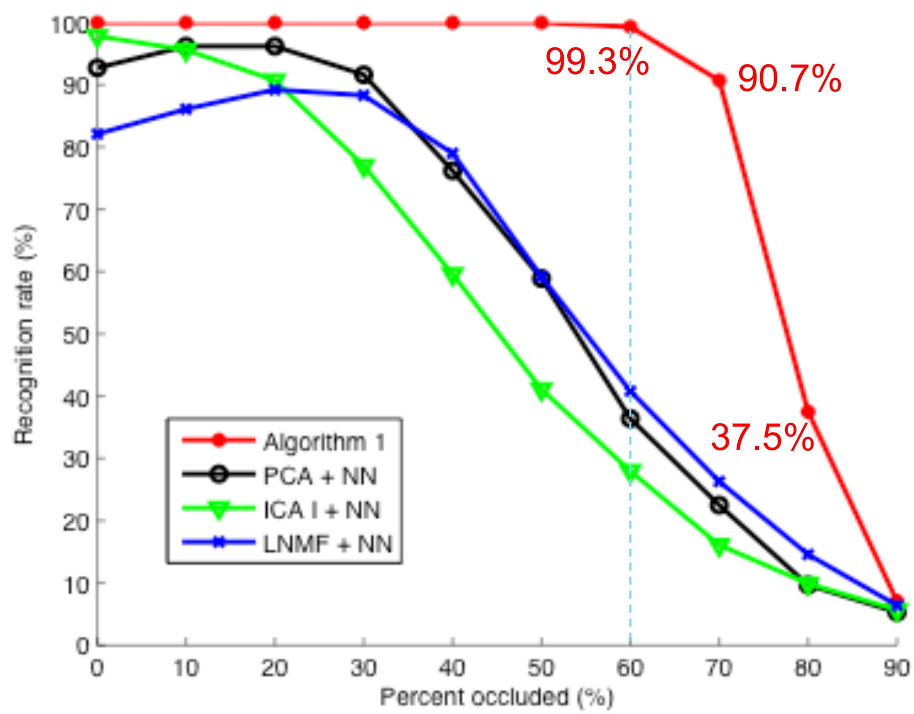
EXPERIMENTS – Varying Level of Random Corruption

Extended Yale B Database
(38 subjects)

Training: subsets 1 and 2 (717 images)
Testing: subset 3 (453 images)

y $\hat{x}_1$ $\hat{e}_1$ $\hat{y}_0 = A\hat{x}_1$

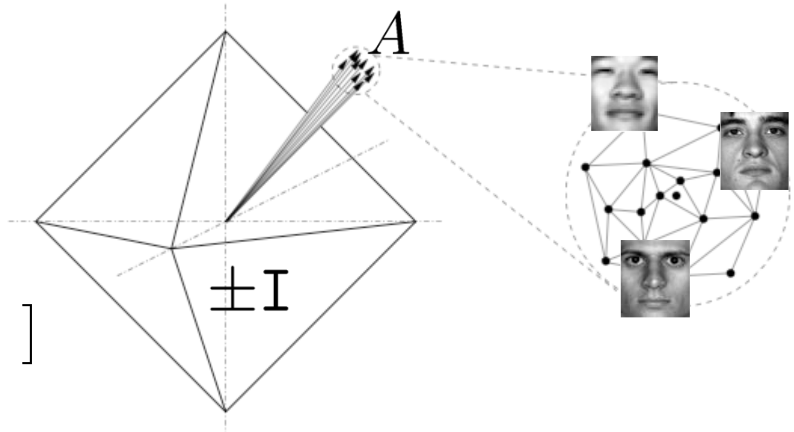




Wright, Yang, Ganesh, Sastry, and Ma. Robust Face Recognition via Sparse Representation, TPAMI 2009

Theory – Geometry and Statistics of Face Images

$$B = [\ A \quad \mathtt{I}\ ]$$



Highly coherent

(volume $\leq 1.5 \times 10^{-229}$)

Theorem 1. *For any $\delta > 0$, $\exists\, \nu_0(\delta) > 0$ such that if $\nu < \nu_0$ and $\rho < 1$, in weak proportional growth, with error support J and signs $\boldsymbol{\sigma}$ chosen uniformly at random,*

$$\lim_{m \to \infty} P_{A,J,\sigma}\left[\, \ell^1\text{-}recoverability\ at\ (I, J, \boldsymbol{\sigma})\ \ \forall\, I \in \binom{[n]}{k_1} \right] \;=\; 1.$$

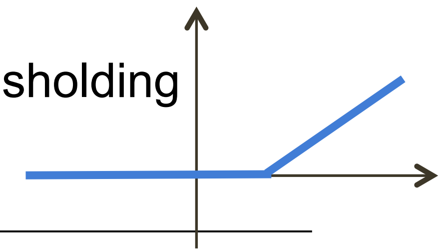
"ℓ^1 recovers any sparse signal from almost any error with density less than 1"

CONTEXT – Basic Algorithm for Sparsity (ISTA)

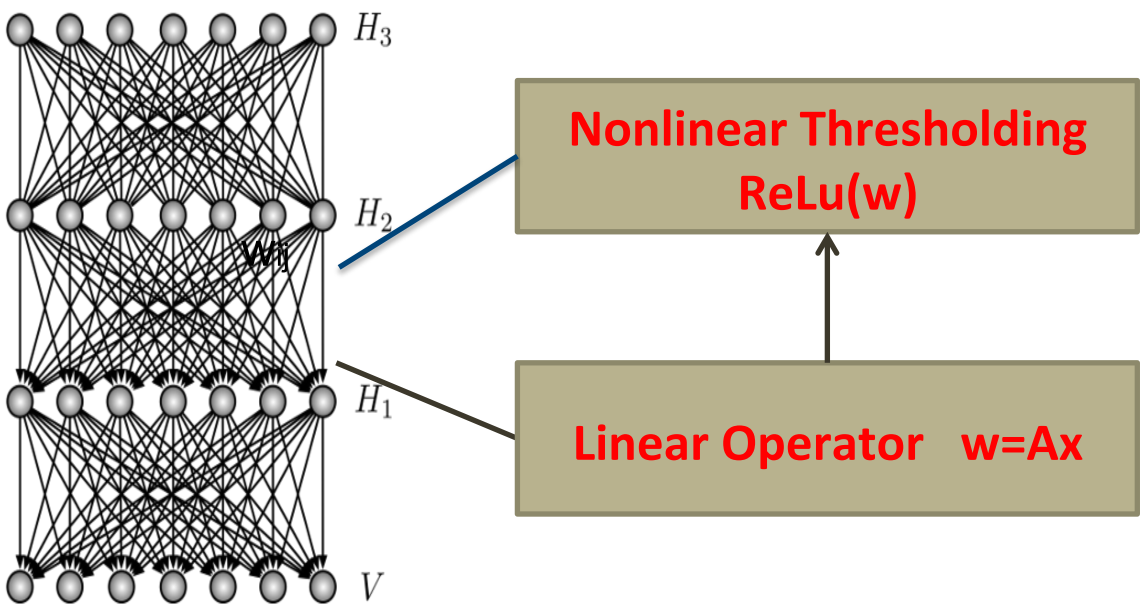
Algorithm 8.1 Iterative Soft-Thresholding Algorithm (ISTA) for BPDN

Problem: $\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1$, given $\boldsymbol{y} \in \mathbb{R}^d$, $\boldsymbol{A} \in \mathbb{R}^{d \times n}$.
Input: $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $L \geq \lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})$.
while $\boldsymbol{x}_k$ not converged $(k = 1, 2, \ldots)$ **do**
    $\boldsymbol{w}_k \leftarrow \boldsymbol{x}_k - \frac{1}{L}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{y})$.
    $\boldsymbol{x}_{k+1} \leftarrow \mathrm{soft}(\boldsymbol{w}_k, \lambda/L)$.
end while
Output: $\boldsymbol{x}_\star \leftarrow \boldsymbol{x}_k$.

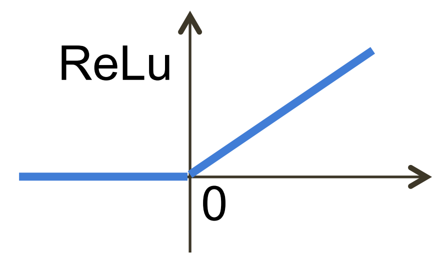
Soft Thresholding



Deep Neural Network Module



ReLu



CONTEXT – Learned ISTA (LISTA)

If only interested in one instance: $y = Ax$ AND with many training data: $\{(y_i, x_i)\}$.
We can optimize the optimization path of ISTA using supervised learning:

Algorithm 3 LISTA::fprop

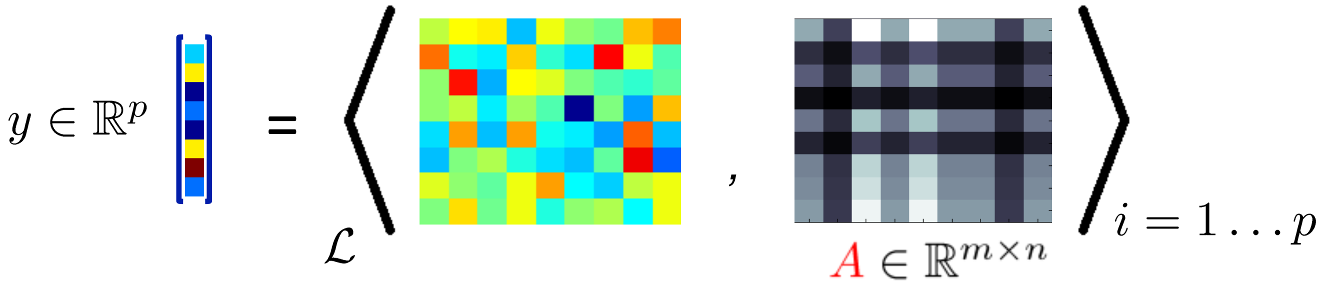
LISTA :: **fprop**(X, Z, W_e, S, θ)
;; Arguments are passed by reference.
;; variables $Z(t)$, $C(t)$ and B are saved for bprop.
$B = W_e X$; $Z(0) = h_\theta(B)$
for $t = 1$ to T **do**
  $C(t) = B + SZ(t-1)$
  $Z(t) = h_\theta(C(t))$
end for
$Z = Z(T)$

Algorithm 4 LISTA::bprop

LISTA :: **bprop**$(Z^*, X, Z, W_e, S, \theta, \delta X, \delta W_e, \delta S, \delta\theta)$
;; Arguments are passed by reference.
;; Variables $Z(t)$, $C(t)$, and B were saved in fprop.
Initialize: $\delta B = 0$; $\delta S = 0$; $\delta\theta = 0$
$\delta Z(T) = (Z(T) - Z^*)$
for $t = T$ down to 1 **do**
  $\delta C(t) = h'_\theta(C(t)).\delta Z(t)$
  $\delta\theta = \delta\theta - \text{sign}(C(t)).\delta C(t)$
  $\delta B = \delta B + \delta C(t)$
  $\delta S = \delta S + \delta C(t) Z(t-1)^T$
  $\delta Z(t-1) = S^T \delta C(t)$
end for
$\delta B = \delta B + h'_\theta(B).\delta Z(0)$
$\delta\theta = \delta\theta - \text{sign}(B).h'_\theta(B)\delta Z(0)$
$\delta W_e = \delta B X^T$; $\delta X = W_e^T \delta B$

CONTEXT – Recent related progress

Low-rank recovery: Given $y = \mathcal{L}[A_0]$, $\mathcal{L} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$, recover A_0.



Impossible in general ($p \ll mn$)

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

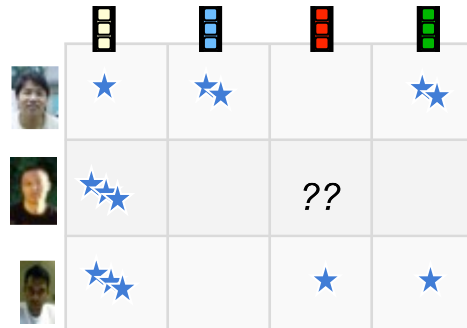
Tractable via convex optimization: $\min \|A\|_*$ s.t. $y = \mathcal{L}(A)$

… 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, Oymak+Tropp'15 …*

CONTEXT – Recent related progress

Matrix completion: Given $y = \mathcal{P}_\Omega[A_0]$, $\Omega \subset [m] \times [n]$, recover A_0.



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

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

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

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

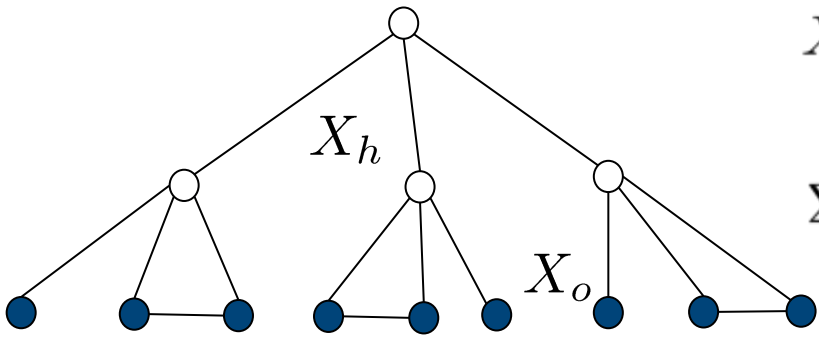
… if Ω is "nice" *(random subset) ...*

… and A_0 interacts "nicely" with $\mathcal{P}_\Omega$ *(A_0 incoherent – not "spiky")*.

Hugely active area: *Candès+Recht '08, Keshevan+Oh+Montonari '09, Candès+Tao '09, Gross '10, Recht '10, Negahban+Wainwright '10, Oymak+Tropp'15…*

CONTEXT – Why Should You Care?

Learning Graphical Models



$$X = (X_o, X_h) \sim \mathcal{N}(0, \Sigma)$$

$$\Sigma = \begin{bmatrix} \Sigma_o & \Sigma_{oh} \\ \Sigma_{ho} & \Sigma_h \end{bmatrix} \Rightarrow \Sigma^{-1} = \begin{bmatrix} J_o & J_{oh} \\ J_{ho} & J_h \end{bmatrix}$$

$X_i,\ X_j$ cond. indep. given other variables $\Leftrightarrow$ $(\Sigma^{-1})_{ij} = 0$

Separation Principle:

$$\begin{array}{ccccc} \Sigma_o^{-1} & = & J_o & - & J_{oh} J_h^{-1} J_{ho} \\ \text{observed} & = & \text{sparse} & + & \text{low-rank} \end{array}$$

- sparse pattern → conditional (in)dependence
- rank of second component → number of hidden variables

CONTEXT – Why Should You Care?

Learning Deep Neural Networks

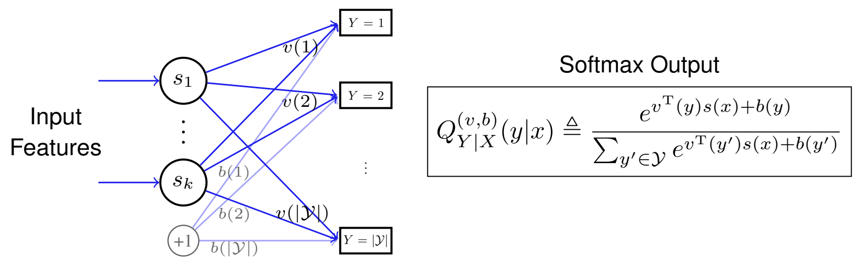


Figure 2. A simple neural network with one layer of hidden nodes, with softmax output, can be viewed as selecting features.

Theorem 1. *The softmax function (17) can be approximated as*

$$Q^{(v,b)}_{Y|X}(y|x) = P_Y(y)\left(1 + \tilde{v}^{\mathrm{T}}(y)s(x) + \tilde{d}(y)\right) + o(\epsilon)$$

and the loss (16), equivalently expressed as the K-L divergence, can be approximated as

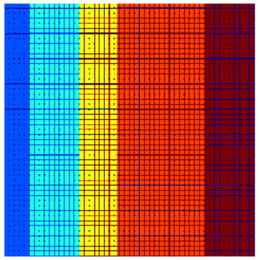
$$D(P_{Y,X}\|P_X\, Q^{(v,b)}_{Y|X}) \qquad (18)$$
$$= \frac{1}{2}\|\tilde{\mathbf{B}} - \Psi\Phi^{\mathrm{T}}\|^2_{\mathrm{F}} + \frac{1}{2}\eta^{(v,b)}(s) + o(\epsilon^2),$$

where $\eta^{(v,b)}(s) \triangleq \mathbb{E}_{P_Y}\left[(\mu_s^{\mathrm{T}}\tilde{v}(Y) + \tilde{d}(Y))^2\right]$. *Moreover, the loss (18) is minimized when* $\tilde{d}(y) + \mu_s^{\mathrm{T}}\tilde{v}(y) = 0$, *and* Φ, Ψ *are designed from*

$$(\Psi, \Phi)^* = \underset{(\Psi,\Phi)}{\arg\min}\, \|\tilde{\mathbf{B}} - \Psi\Phi^{\mathrm{T}}\|^2_{\mathrm{F}}. \qquad (19)$$

From information-theoretic perspective, DNNs (with softmax objective) is to learn a low-rank approximation of the joint distribution *P(X,Y)* of the input *X* and output *Y*.

CONTEXT – Low-dimensional Models



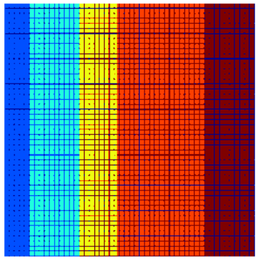
*The data should be **low-dimensional (low-rank)**:*

$$A = [\mathbf{a}_1 \mid \cdots \mid \mathbf{a}_n] \in \mathbb{R}^{m \times n}, \qquad \mathrm{rank}(A) \ll m.$$

CONTEXT – Low-dimensional Models

*The data should be **low-dimensional**:*

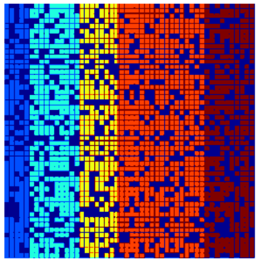
$$A = [\mathbf{a}_1 \mid \cdots \mid \mathbf{a}_n] \in \mathbb{R}^{m \times n}, \qquad \mathrm{rank}(A) \ll m.$$



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

$$A + E, \qquad |E_{ij}|$$

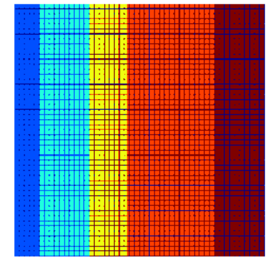
E_{ij} arbitrarily large, but most are zero.



CONTEXT – Low-dimensional Models

*The data should be **low-dimensional**:*

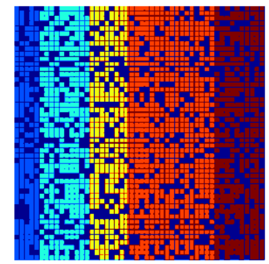
$$A = [\mathbf{a}_1 \mid \cdots \mid \mathbf{a}_n] \in \mathbb{R}^{m \times n}, \qquad \mathrm{rank}(A) \ll m.$$



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

$$A + E, \qquad |E_{ij}|$$

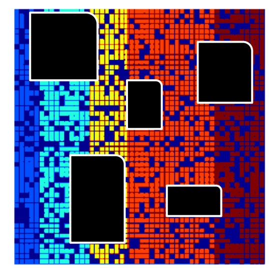
$$E_{ij} \text{ arbitrarily large, but most are zero.}$$



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

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

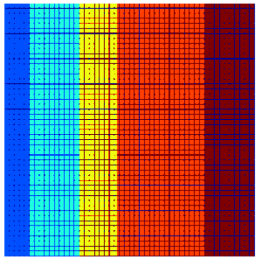
$$\Omega \subset [m] \times [n] \text{ the set of observed entries.}$$



CONTEXT – Low-dimensional Models

*The data should be **low-dimensional**:*

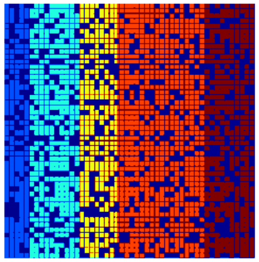
$$A = [\mathbf{a}_1 \mid \cdots \mid \mathbf{a}_n] \in \mathbb{R}^{m \times n}, \qquad \operatorname{rank}(A) \ll m.$$



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

$$A + E, \qquad |E_{ij}|$$

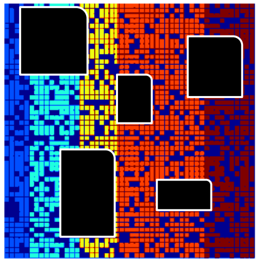
E_{ij} arbitrarily large, but most are zero.



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

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

$\Omega \subset [m] \times [n]$ the set of observed entries.



… special cases of a more general problem:

$$D = \mathcal{L}_1(A) + \mathcal{L}_2(E) + Z \qquad A, E \text{ either sparse or low-rank}$$

THIS TALK

Given observations $D = \mathcal{P}_Q[A + E + Z]$, *with*

A *low-rank,*

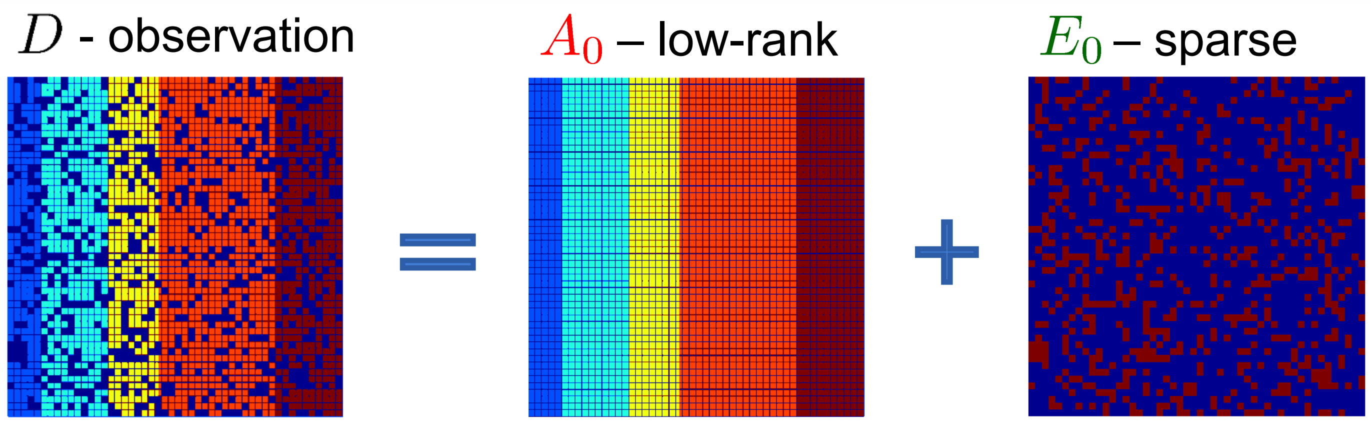
E *sparse,*

Z *small, dense noise,*

recover a good estimate of A *and* E.

- **Theory and Algorithm**
  - **Provably Correct and Tractable Solution**
  - **Provably Optimal and Efficient Algorithms**
- **Potential Applications**
  - **Visual Data (Restoration, Reconstruction, Recognition)**
  - **Other Data**
- **Extensions and Conclusions**

ROBUST PCA – Problem Formulation



Problem: Given $D = A_0 + E_0$, recover A_0 and E_0 .

Low-rank component **Sparse component (gross errors)**

Numerous approaches in the literature:

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

Key question: ***can guarantee correctness with an efficient algorithm?***

ROBUST PCA – Convex Surrogates for Sparsity and Rank

Seek the lowest-rank A that agrees with the data up to some sparse error E:

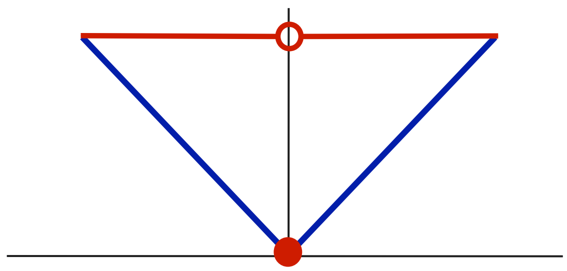
$$\min\ \operatorname{rank}(A) + \gamma \|E\|_0 \quad \text{subj} \quad A + E = D.$$

But INTRACTABLE! Relax with convex surrogates:

$$\|E\|_0 = \#\{E_{ij} \neq 0\} \quad \to \quad \|E\|_1 = \textstyle\sum_{ij} |E_{ij}|. \qquad \text{L}_1 \text{ norm}$$

$$\operatorname{rank}(A) = \#\{\sigma_i(A) \neq 0\} \quad \to \quad \|A\|_* = \textstyle\sum_i \sigma_i(A). \qquad \text{Nuclear norm}$$

Convex envelope over $B_{2,2} \times B_{1,\infty}$



ROBUST PCA – By Convex Optimization

Seek the lowest-rank A that agrees with the data up to some sparse error E:

$$\min \ \operatorname{rank}(A) + \gamma \|E\|_0 \quad \text{subj} \quad A + E = D.$$

But INTRACTABLE! Relax with convex surrogates:

$$\|E\|_0 = \#\{E_{ij} \neq 0\} \quad \rightarrow \quad \|E\|_1 = \textstyle\sum_{ij} |E_{ij}|. \qquad \text{L}_1 \text{ norm}$$

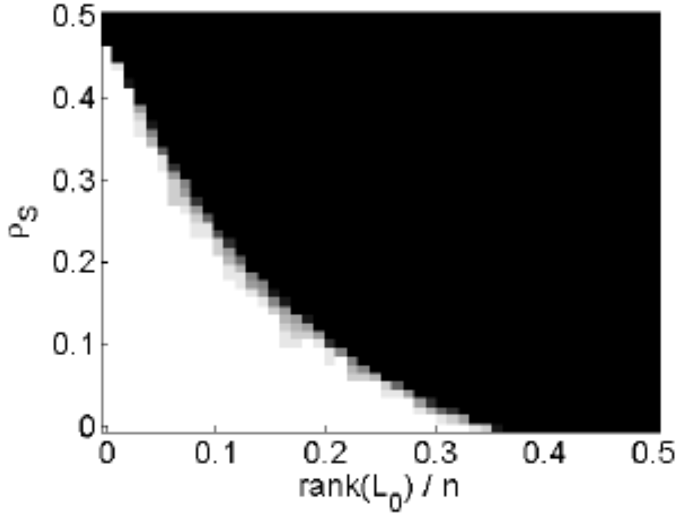
$$\operatorname{rank}(A) = \#\{\sigma_i(A) \neq 0\} \quad \rightarrow \quad \|A\|_* = \textstyle\sum_i \sigma_i(A). \qquad \text{Nuclear norm}$$

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

Semidefinite program, solvable in polynomial time

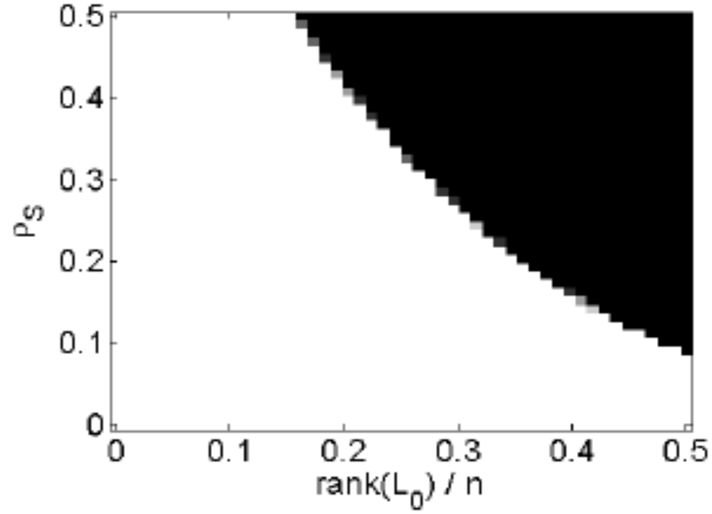
ROBUST PCA – When the Convex Program Works?

$D = A + E$



Robust PCA, Random Signs

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



Matrix Completion

White regions are instances with perfect recovery.

Correct recovery when A is indeed **low-rank** and E is indeed **sparse**?

MAIN THEORY – Exact Solution by Convex Optimization

Theorem 1 (Principal Component Pursuit). *If* $A_0 \in \mathbb{R}^{m \times n}$, $m \geq n$ *has rank*

m

Non-adaptive weight factor

and E_0 *has Bernoulli support with error probability* $\rho \leq \rho_s^\star$, *then with very high probability*

$$(A_0, E_0) \;=\; \arg\min \|A\|_* + \frac{1}{\sqrt{m}}\|E\|_1 \quad \text{subj} \quad A + E = A_0 + E_0,$$

and the minimizer is unique.

GREAT NEWS: *"Convex optimization recovers almost any matrix of rank* $O\left(\frac{m}{\log^2 n}\right)$ *from errors corrupting* $O(mn)$ *of the observations!"*

Candes, Li, Ma, and Wright, Journal of the ACM, May 2011.

MAIN THEORY – Corrupted, Incomplete Matrix

$$D = \mathcal{P}_\Omega[\ A_0\ +\ E_0\ ], \qquad \Omega \sim \mathrm{uni}\binom{[m]\times[n]}{mn}$$

Theorem 2 (Matrix Completion and Recovery). *If* $A_0, E_0 \in \mathbb{R}^{m\times n}, m \geq n$*, with*

$$\mathrm{rank}(A_0)\ \leq\ C\,\frac{n}{\mu\log^2(m)}, \quad \textit{and} \quad \|E_0\|_0\ \leq\ \rho^\star mn,$$

and we observe only a random subset of size

$$|\Omega| = mn/10$$

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

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

uniquely recovers (A_0, E_0).

MAIN THEORY – With Dense Errors and Noise

Theorem 3 (Dense Error Correction). *If A_0 has rank $r \leq \rho_r \frac{m}{\mu^2 \log^2(n)}$ 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.

Theorem 4 (Robust PCA with Noise). *Given $D = A_0 + E_0 + Z$ for any $\|Z\|_F \leq \eta$, if A_0 has rank $r \leq \rho_r \frac{m}{\mu^2 \log^2(n)}$ and E_0 has Bernoulli support with error probability $\rho \leq \rho_s^\star$, then with very high probability*

$$(\hat{A}, \hat{E}) \;=\; \arg\min \|A\|_* + \frac{1}{\sqrt{m}} \|E\|_1 \quad \text{subj} \quad \|D - A - E\| \leq \eta,$$

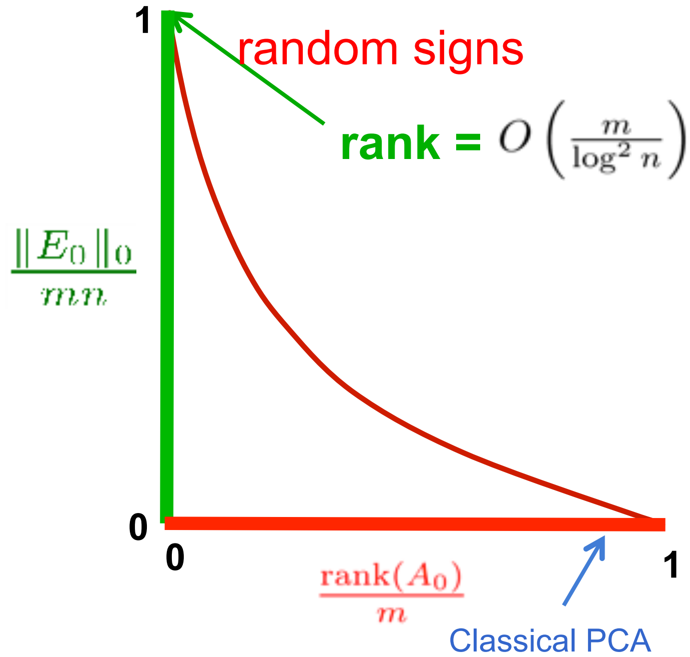
sastisfies $\|(\hat{A}, \hat{E}) - (A_0, E_0)\| \leq C\eta$ for some constant $C > 0$.

BIG PICTURE – Landscape of Theoretical Guarantees

Many have made contributions in the past few years:

Matrix Recovery (RPCA)

$$D = A + E$$

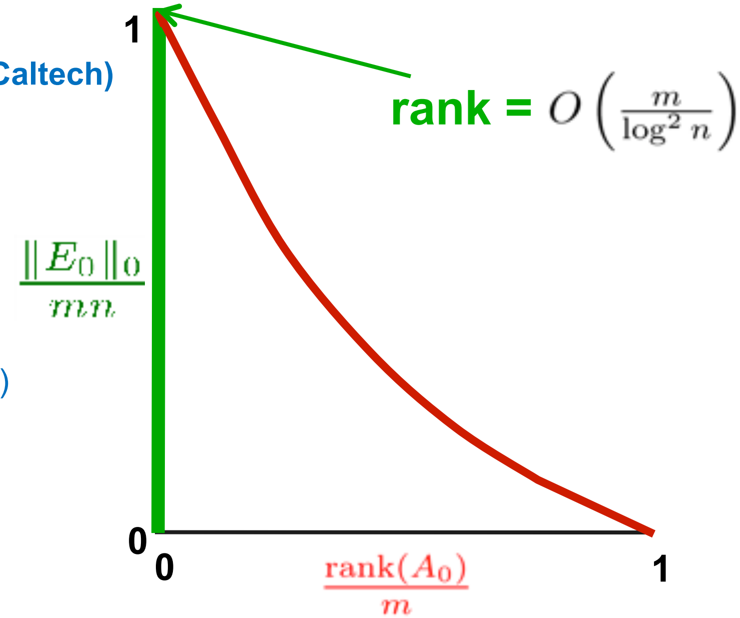


D. Gross
E. Candes (Stanford)
B. Recht (UC Berkeley)
J. Wright (Columbia)
J. Tropp (Caltech)
V. Chandrasekharan (Caltech)

B. Hassibi (Caltech)
P. Parrilo (MIT)
A. Willsky (MIT)
B. Hastie (Stanford)
C. Montanari (Stanford)
M. Jordan (Berkeley)
M. Wainwright (Berkeley)
B. Yu (Berkeley)
A. Singer (Princeton)
T. Tao (UCLA)
S. Osher (UCLA)
O. Milenkovic (UIUC)
Y. Bresler (UIUC)
Y. Ma (UIUC)
M. Fazel (U Wash.)
… …

Matrix Completion

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

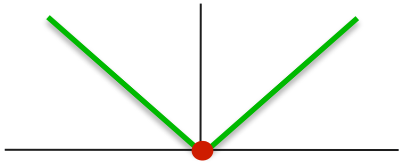


Universality of phase transition (Oymak & Tropp). But does not yet apply here…

ALGORITHMS – Are scalable solutions possible?

Seemingly BAD NEWS: Our optimization problem

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

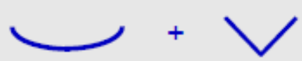


is high-dimensional and non-smooth.

Convergence rate of solving a generic convex program: $\min_x f(x)$

Second-order Newton method, linear rate of convergence , but not scalable!

First-order methods depend strongly on the smoothness of f:

Function class $\mathcal{F}$	Suboptimality $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)$
smooth  f convex, differentiable $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}')\| \le L \|\boldsymbol{x} - \boldsymbol{x}'\|$	$\frac{CL\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{k^2} = \Theta\left(\frac{1}{k^2}\right)$
smooth + structured nonsmooth:  $F = f + g$ f, g convex, $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}')\| \le L \|\boldsymbol{x} - \boldsymbol{x}'\|$	$\frac{CL\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{k^2} = \Theta\left(\frac{1}{k^2}\right)$
nonsmooth  f convex $\|f(\boldsymbol{x}) - f(\boldsymbol{x}')\| \le M \|\boldsymbol{x} - \boldsymbol{x}'\|$	$\frac{CM\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|}{\sqrt{k}} = \Theta\left(\frac{1}{\sqrt{k}}\right)$

Y. Nesterov, Introductory Lectures on Convex Optimization: A Basic Course, 2003.

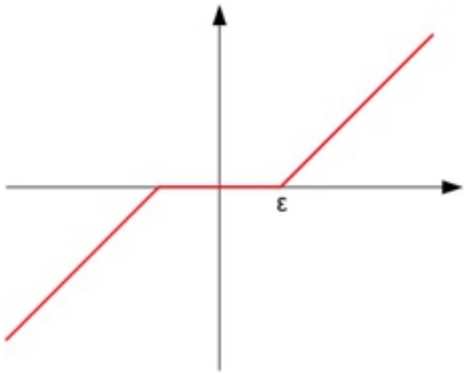
ALGORITHMS – Why are scalable solutions possible?

GOOD NEWS: The objective function has special structures

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

KEY OBSERVATION: Simple solutions for the proximal operations, given by soft-thresholding the entries or singular values of the matrix, respectively.

$$\mathcal{S}_\varepsilon(Q) = \operatorname{argmin}_X \varepsilon\|X\|_1 + \frac{1}{2}\|X - Q\|_F^2$$
$$\mathcal{D}_\varepsilon(Q) = \operatorname{argmin}_X \varepsilon\|X\|_* + \frac{1}{2}\|X - Q\|_F^2$$



For composite functions $F = f + g$, with f smooth,
***if g has an efficient proximal operator**, we achieve*
the same (optimal) rate as if F was smooth.

ALGORITHMS – Evolution of scalable algorithms

GOOD NEWS: Scalable first-order gradient-descent algorithms:

- Proximal Gradient [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].

For a 1000x1000 matrix of rank 50, with 10% (100,000) entries randomly corrupted: $\min \ \|A\|_* + \lambda\|E\|_1 \ \ \text{subj} \ \ A + E = D.$

Algorithms	Accuracy	Rank	||E||_0	# iterations	time (sec)
IT	5.99e-006	50	101,268	8,550	119,370.3
DUAL	8.65e-006	50	100,024	822	1,855.4
APG	5.85e-006	50	100,347	134	1,468.9
APG_P	5.91e-006	50	100,347	134	82.7
$EALM_P$	2.07e-007	50	100,014	34	37.5
$IALM_P$	3.83e-007	50	99,996	23	11.8

10,000 times speedup!

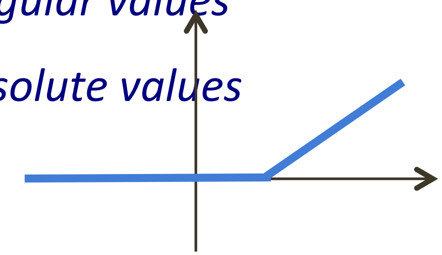
Lin, Chen, and Ma, UILU-ENG-09-2214, 2010.

ALGORITHMS – Evolution of scalable algorithms

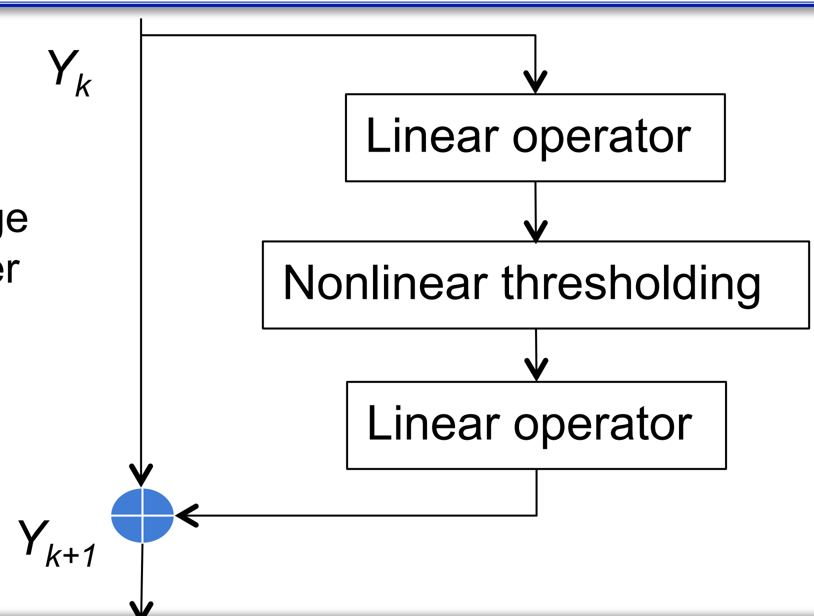
A scalable algorithm: alternating direction method (ADMoM) for ALM:

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

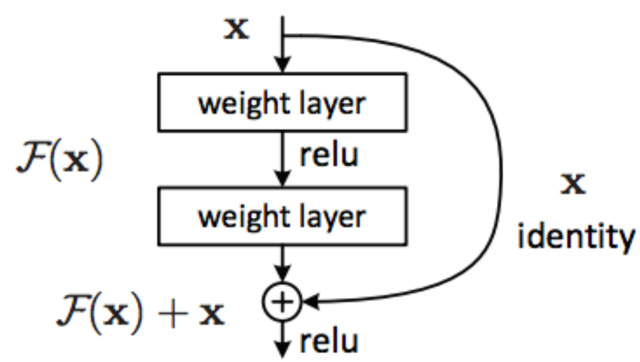
$$\text{repeat}\begin{cases} A_{k+1} = \mathcal{D}_{\mu_k^{-1}}(D - E_k + Y_k/\mu_k), & \textit{Shrink singular values} \\ E_{k+1} = \mathcal{S}_{\lambda\mu_k^{-1}}(D - A_{k+1} + Y_k/\mu_k), & \textit{Shrink absolute values} \\ Y_{k+1} = Y_k + \mu_k(D - A_{k+1} - E_{k+1}). \end{cases}$$



Lagrange Multiplier Update



Residual Network Module



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: $\min \|A\|_* + \lambda\|E\|_1 \;\; \text{subj} \;\; A + E = D.$

Algorithms	Accuracy	Rank	||E||_0	# iterations	time (sec)
IT	5.99e-006	50	101,268	8,550	119,370.3
DUAL	8.65e-006	50	100,024	822	1,855.4
APG	5.85e-006	50	100,347	134	1,468.9
APG_P	5.91e-006	50	100,347	134	82.7
$EALM_P$	2.07e-007	50	100,014	34	37.5
$IALM_P$	3.83e-007	50	99,996	23	11.8

10,000 times speedup!

***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!

f restricted strong convex: $O(\log(1/\varepsilon))$ [Agarwal, Negahban, Wainwright, NIPS 2010]
f smooth, ∇f Lipschitz: $O(\varepsilon^{-1/2})$
f differentiable: $O(\varepsilon^{-1})$
f non-smooth: $O(\varepsilon^{-2})$

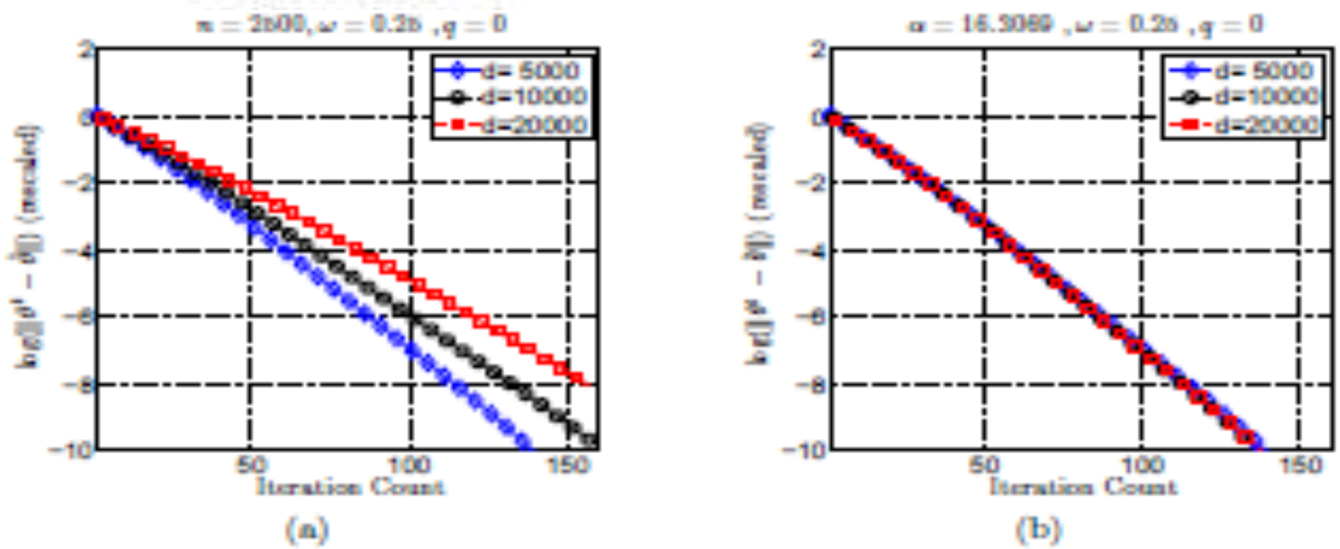


Figure 1. Convergence rates of projected gradient descent in application to Lasso programs (ℓ_1-constrained least-squares). Each panel shows the log optimization error $\log\|\theta^t - \widehat{\theta}\|$ versus the iteration number t. Panel (a) shows three curves, corresponding to dimensions $d \in \{5000, 10000, 20000\}$, sparsity $s = \lceil\sqrt{d}\rceil$, 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 ($\alpha = \frac{n}{s \log d}$), all three convergence rates should be roughly the same, as predicted by the theory.

ALGORITHMS – Recap and Conclusions

Key challenges of **nonsmoothness** and **scale** can be mitigated by using **special structure** in sparse and low-rank optimization problems:

Efficient proximity operators $\Rightarrow$ *proximal gradient methods*

Separable objectives $\Rightarrow$ *alternating directions methods*

Efficient **moderate-accuracy solutions** for **very large problems**.

Special tricks can further improve specific cases (factorization for low-rank)

Techniques in this literature apply quite broadly.

Extremely useful tools for creative problem formulation / solution.

Fundamental **theory** guiding engineering **practice**:

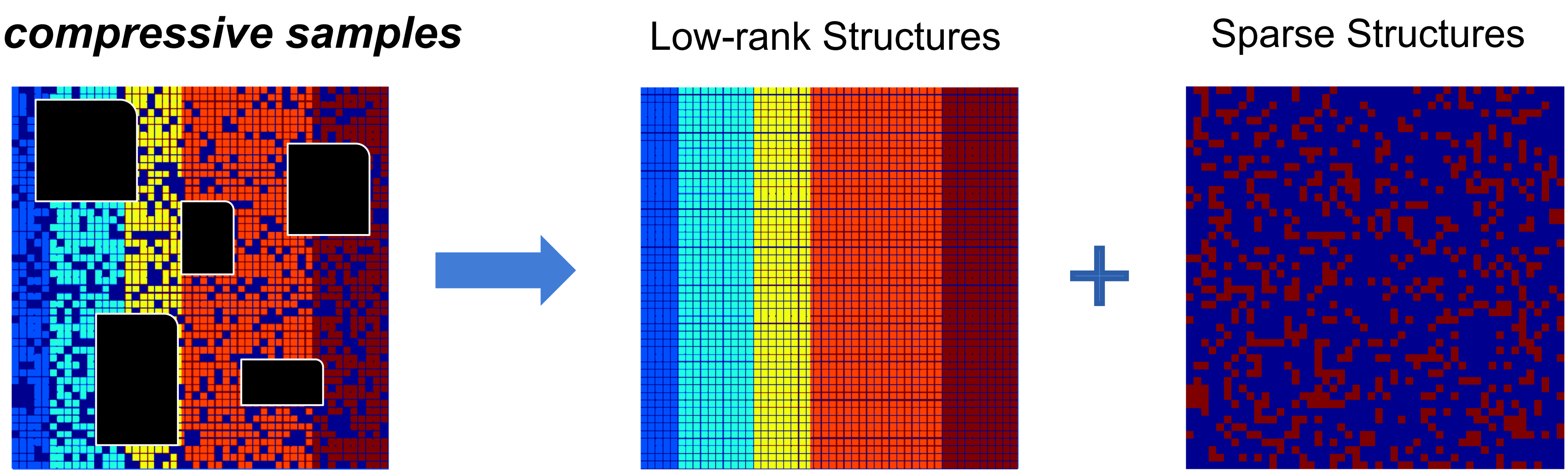
What are the basic principles and limitations?
What specific structure in my problem can allow me to do better?

APPLICATIONS

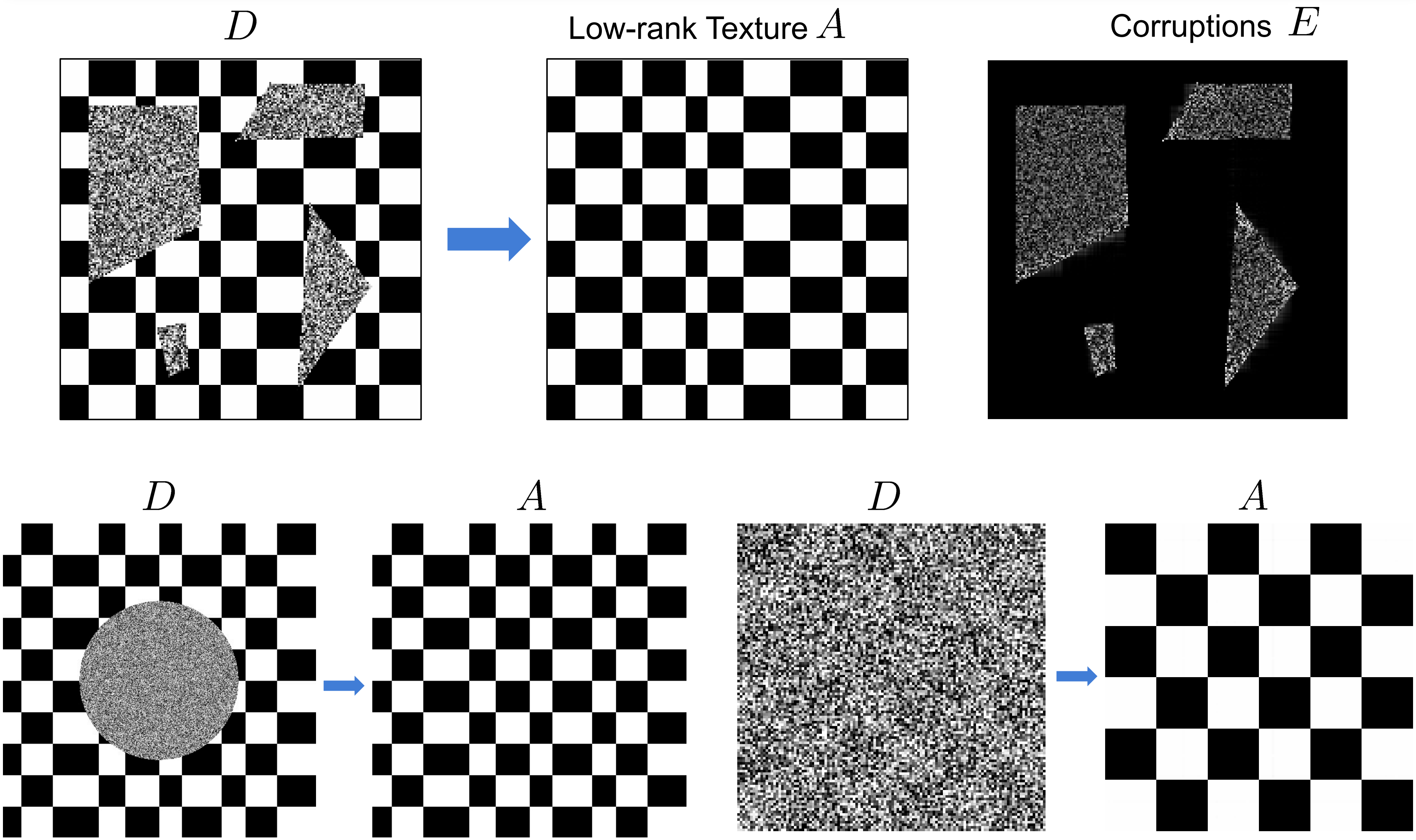
- **Repairing Images and Videos**
  - Image Repairing, Background Extraction, Street Panorama
- **Reconstructing 3D Geometry**
  - Shape from Texture, Featureless 3D Reconstruction
- **Registering Multiple Images**
  - Multiple Image Alignment, Video Stabilization
- **Recognizing Objects**
  - Faces, 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.



Repairing Images – Highly Robust Repairing of Low-rank Textures!



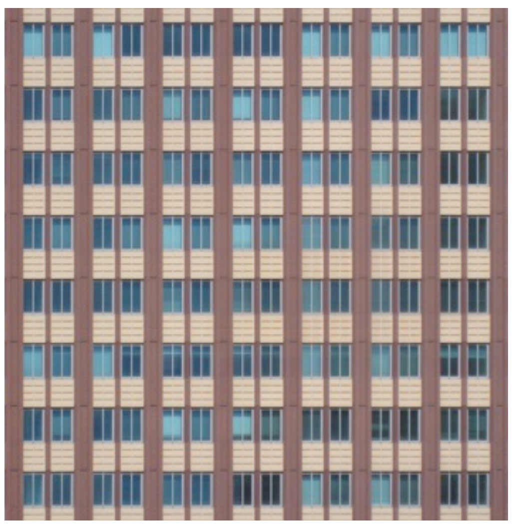
Liang, Ren, Zhang, and Ma, in ECCV 2012.

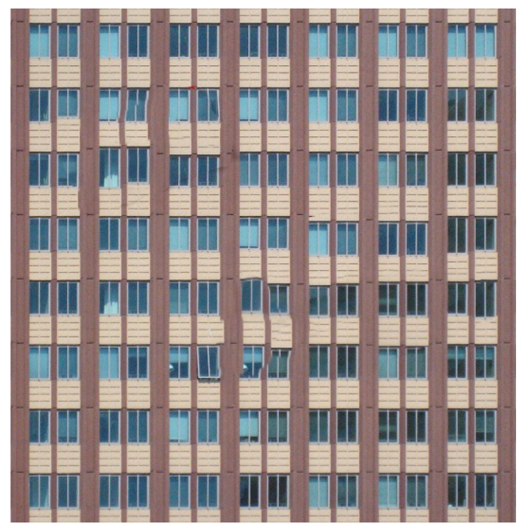
Repairing Low-rank Textures

Low-rank Method | Photoshop

Input

Output

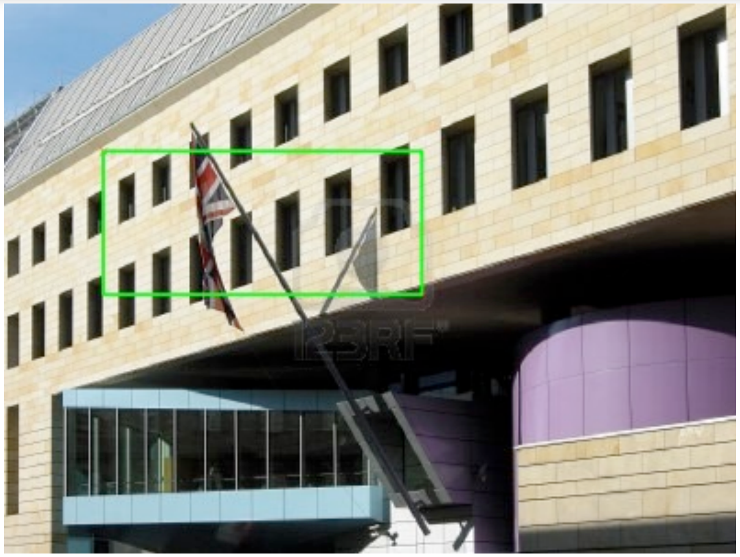




Repairing (Distorted) Low-rank Textures

	Low-rank Method	Photoshop
Input		
Output		

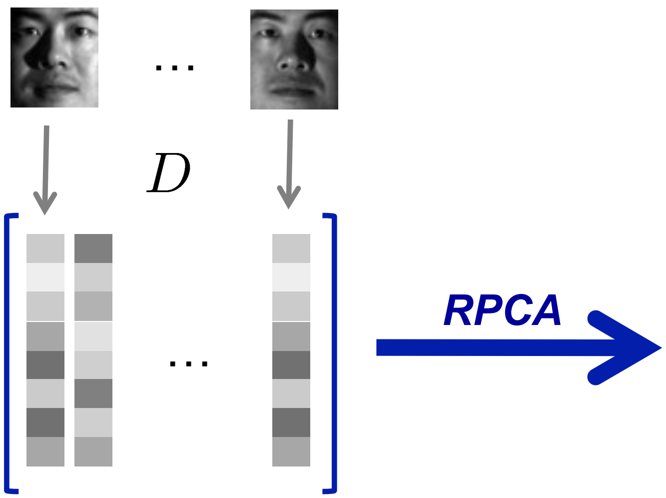
Structured Texture Completion and Repairing

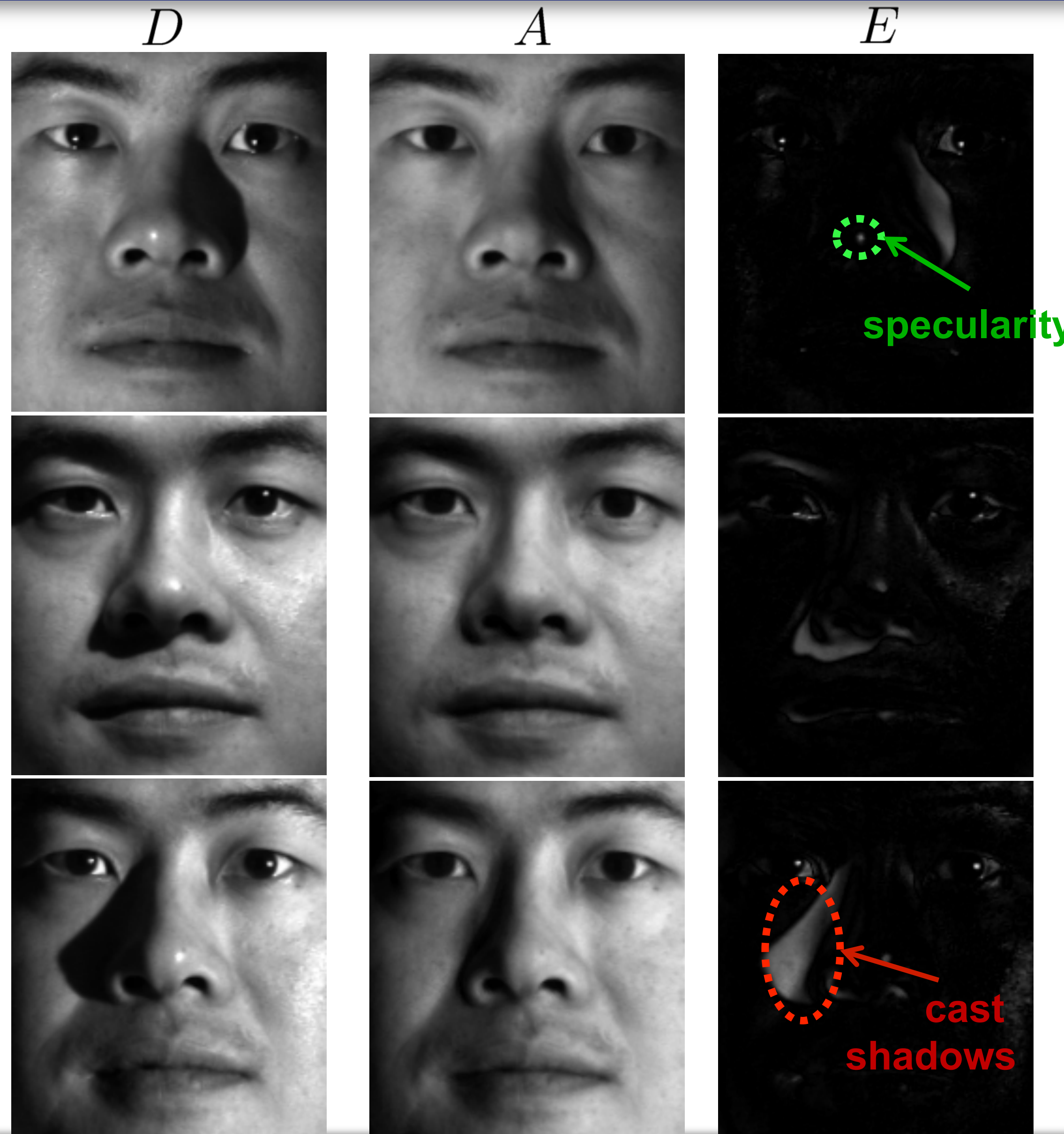


Liang, Ren, Zhang, and Ma, in ECCV 2012.

Repairing Multiple Correlated Images

58 images of one person under varying lighting:





Repairing Images – robust photometric stereo

Input images

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

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

(a) Ground truth (b) Our method (c) Least Squares (d) Error map (our method) (e) Error map (LS)

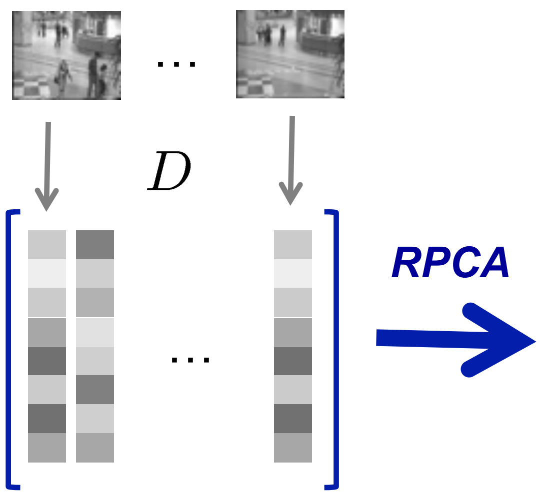
Mean error	**0.014°**	0.96°
Max error	**0.20°**	8.0°

Repairing Video Frames – background modeling from video

Surveillance video

200 frames,
144 x 172 pixels,

Significant foreground motion

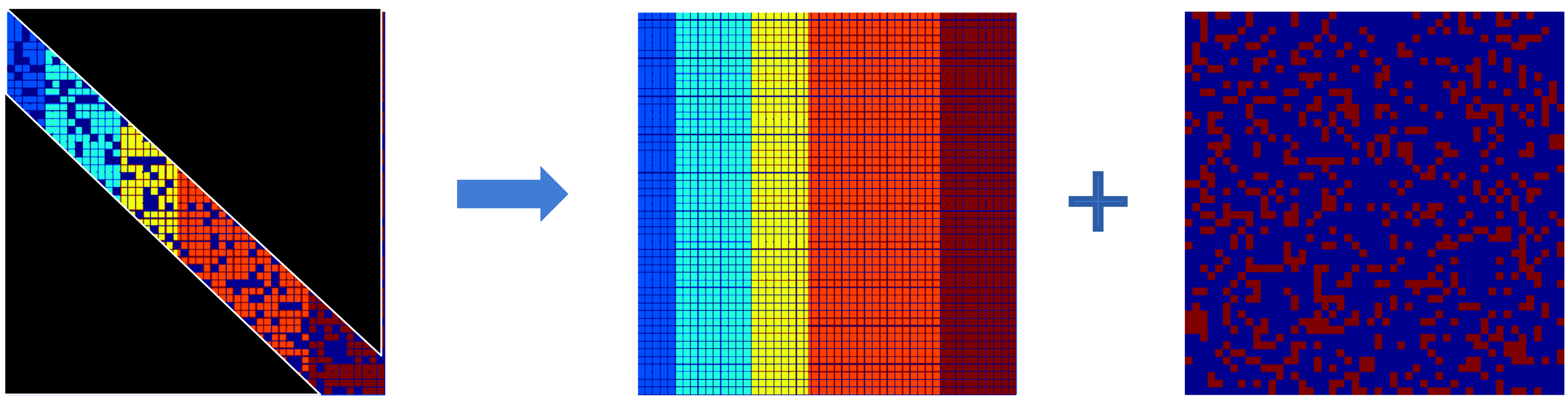


Video D = Low-rank appx. A + Sparse error E

Implications – Highly Compressive Sensing of Structured Information!

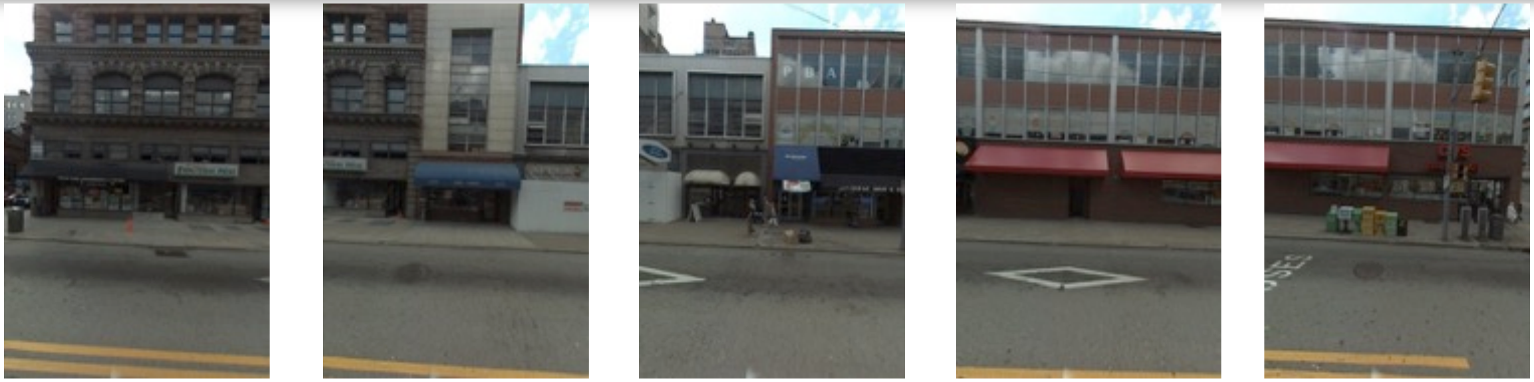
Recover low-dimensional structures from diminishing fraction of corrupted measurements.

compressive samples Low-rank Structures Sparse Structures



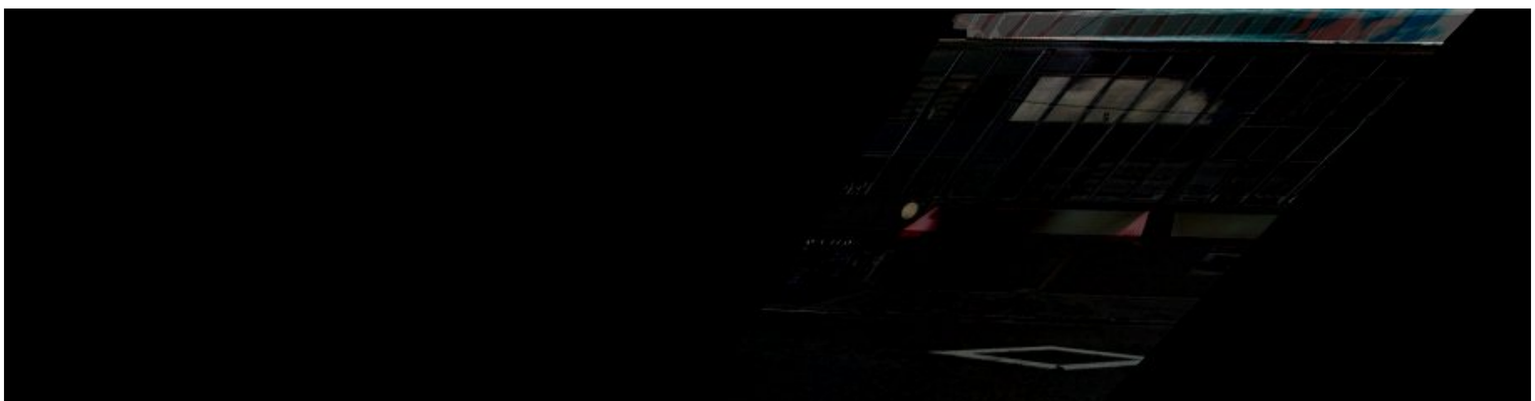
Repairing Video Frames – Street Panorama

D



A

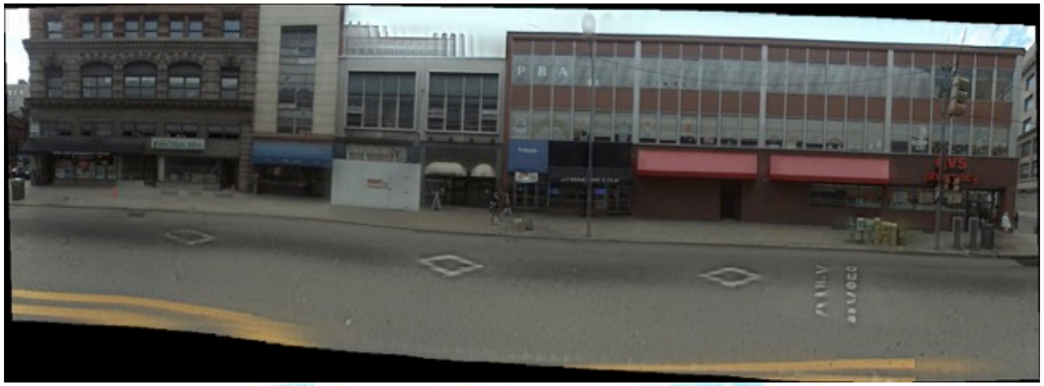
E



Repairing Video Frames – Street Panorama

Low-rank

AutoStitch



Photoshop

Repairing Video Frames – Street Panorama

Low-rank

AutoStitch

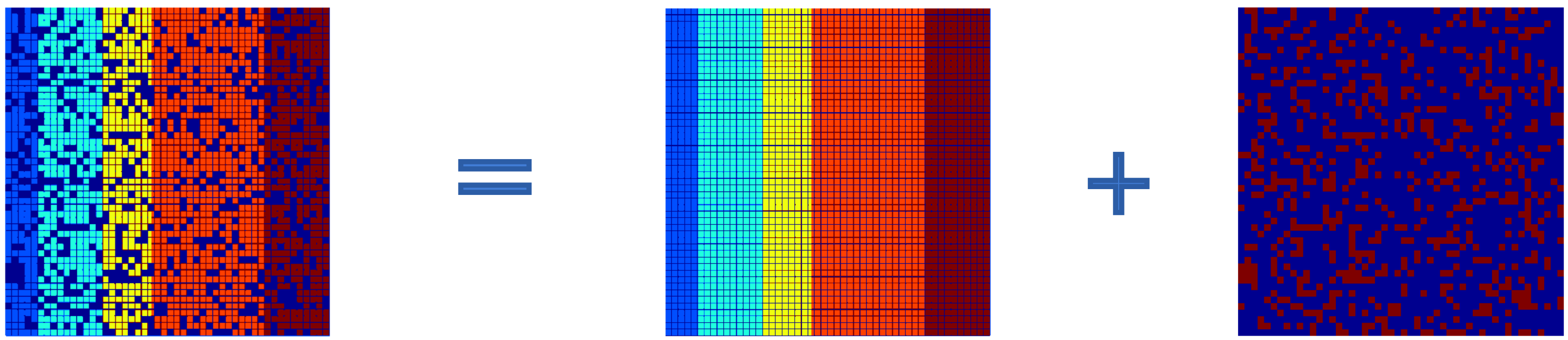
Photoshop

Sensing or Imaging of Low-rank and Sparse Structures

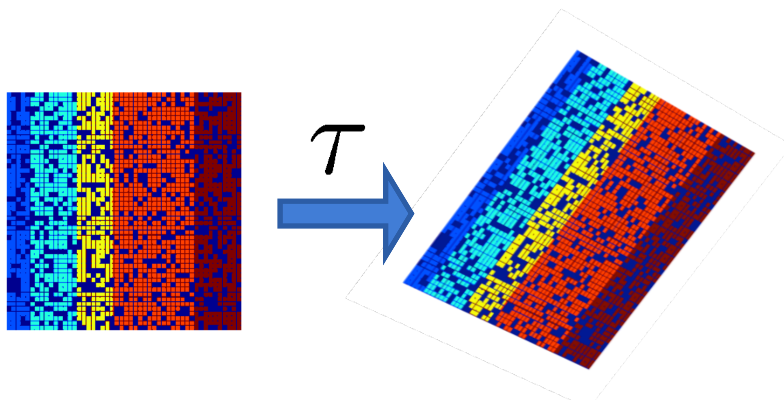
Fundamental Problem: ***How to recover low-rank and sparse structures from corrupted data***

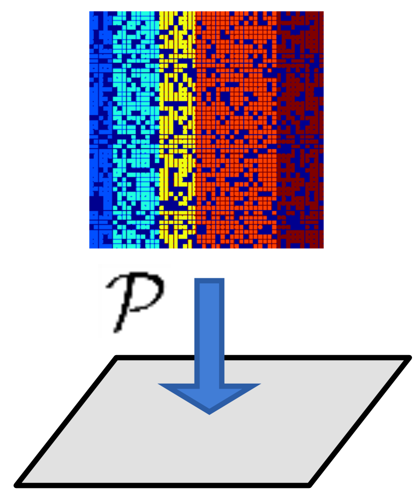
Low-rank Structures

Sparse Structures



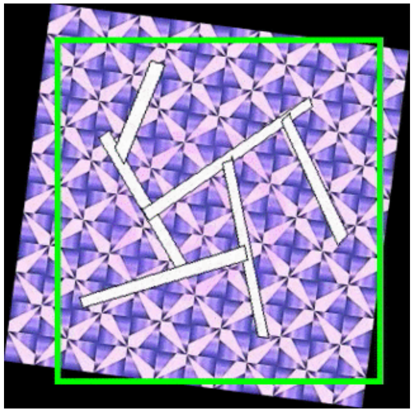
subject to either nonlinear deformation τ or linear compressive sampling $\mathcal{P}$?





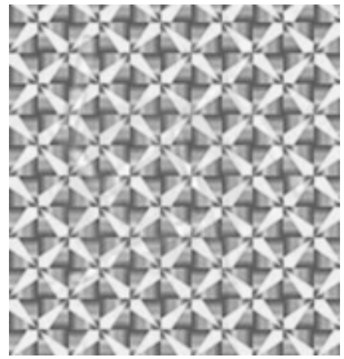
Reconstructing 3D Geometry and Structures

D – deformed observation



$\circ\,\tau \;=$

A – low-rank structures



$+$

E – sparse errors

Problem: Given $D \circ \tau = A_0 + E_0$, recover τ , A_0 and E_0 simultaneously.

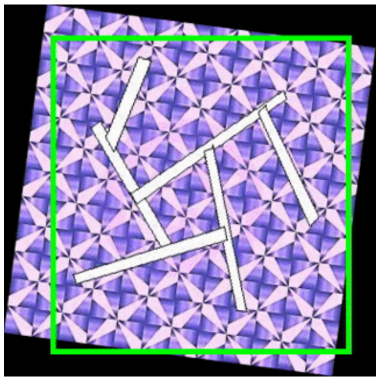
Low-rank component
(regular patterns…)

Sparse component
(occlusion, corruption, foreground…)

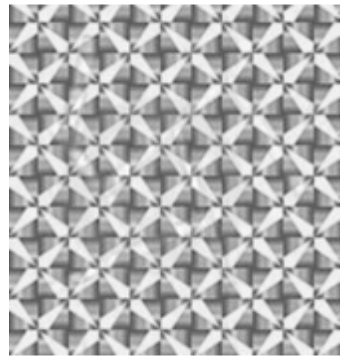
Parametric deformations
(affine, projective, radial distortion, 3D shape…)

Transform Invariant Low-rank Textures (TILT)

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



$\circ \, \tau \; =$



$+$

Objective: *Transformed Principal Component Pursuit:*

$$\min \; \|A\|_* + \lambda \|E\|_1 \quad \text{subj} \quad A + E = D \circ \tau$$

Solution: *Iteratively solving the linearized convex program:*

$$\min \; \|A\|_* + \lambda \|E\|_1 \quad \text{subj} \quad A + E = D \circ \tau_k + J \cdot \Delta\tau$$

Or reduced version: $\quad \text{subj} \quad \mathcal{P}_Q[A + E] = \mathcal{P}_Q[D \circ \tau_k], \; \mathcal{P}_Q[J] = 0$

THEORY – Compressive Robust PCA

Theorem 5 (Compressive Principal Component Pursuit). *Let* $A_0 \in \mathbb{R}^{m\times n}$, $m \geq n$ *have rank* $r \leq \rho_r \frac{m}{\mu^2 \log^2(n)}$, *and* E_0 *have a Bernoulli support with error probability* $\rho < \rho^\star$. *Let* $Q^\perp$ *be a random subspac of* $\mathbb{R}^{m\times n}$ *of dimension*

$$\dim(Q) \geq C_Q(\rho mn + mr) \cdot \log^2 m,$$

distributed according to the Haar measure, independent of the support of E_0. *Then with very high probability*

$$(A_0, E_0) \;=\; \arg\min \|A\|_* + \frac{1}{\sqrt{m}}\|E\|_1 \quad \text{subj} \quad \mathcal{P}_Q[A+E] = \mathcal{P}_Q[A_0+E_0],$$

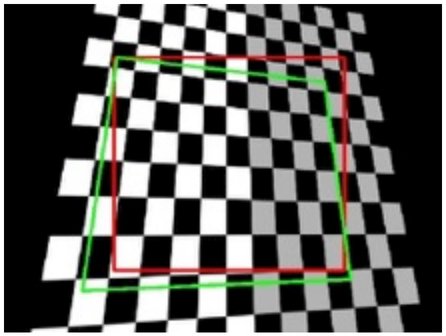
for some numerical constant ρ_r, C_p *and* $\rho^\star$, *and the minimizer is unique.*

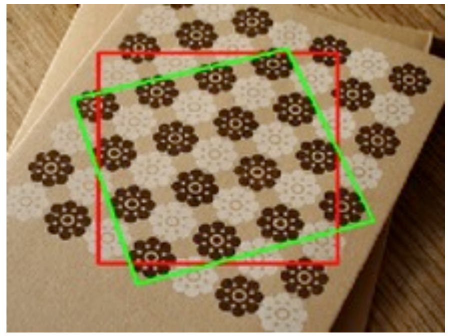
A nearly optimal lower bound on minimum # of measurements!

Wright, Ganesh, Min, and Ma, IMA Information & Inference 2015, the Best Paper 2nd Prize

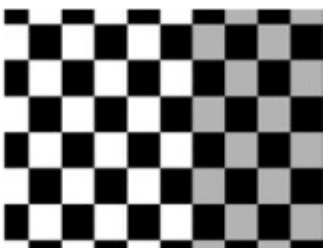
TILT – Shape from texture

Input (red window D)



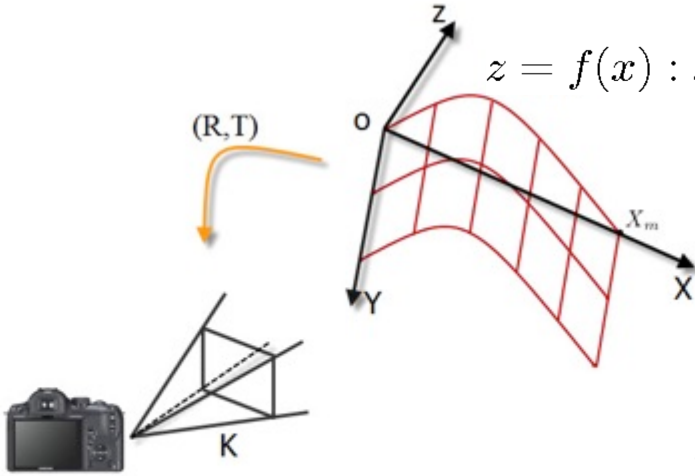


Output (rectified green window A)



Zhang, Liang, Ganesh, Ma, ACCV’10, IJCV’12

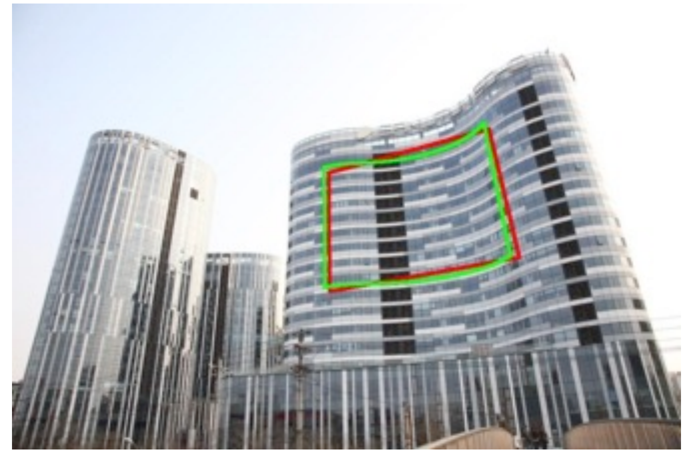
TILT – Shape and geometry from textures



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

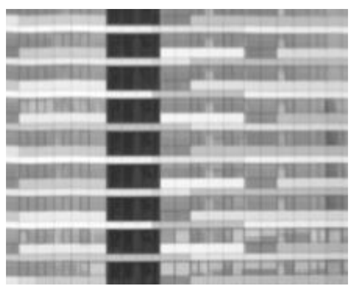
$$D \circ \tau = A + E$$

$$\tau = (K, R, T, \{a_i\})$$

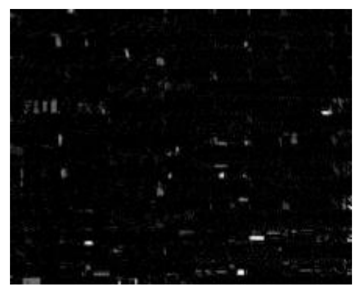


D

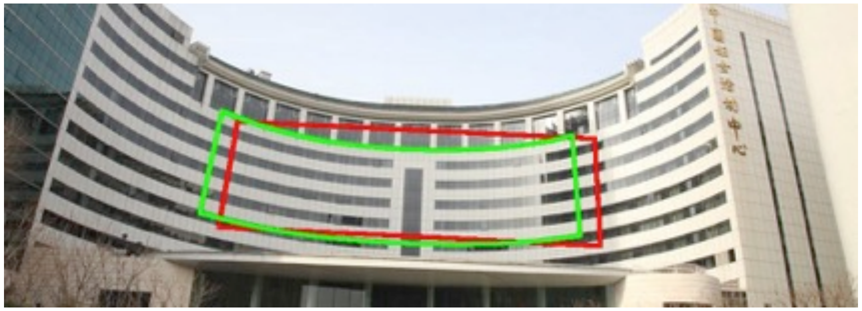
A



E

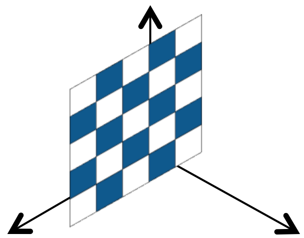


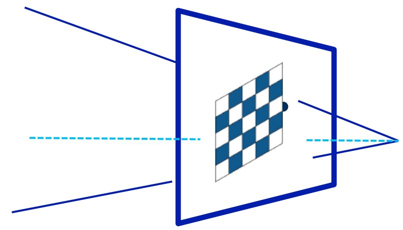
TILT – Virtual reality



Zhang, Liang, and Ma, in ICCV 2011

TILT – Camera Calibration with Radial Distortion

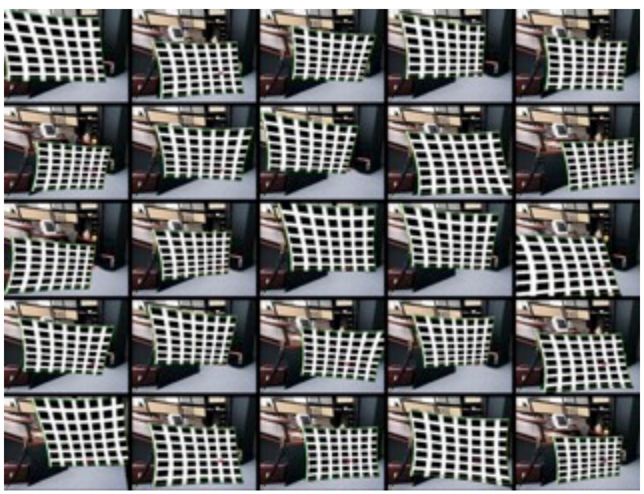


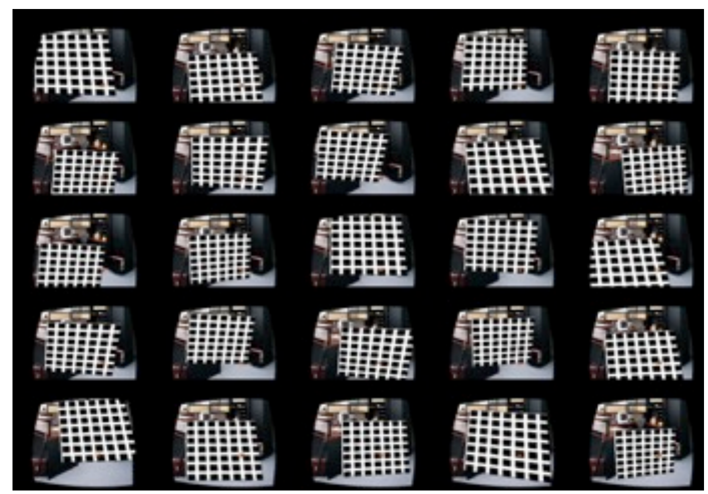


$$r = \sqrt{x_0^2 + y_0^2}, f(r) = 1 + kc(1)r^2 + kc(2)r^4 + kc(5)r^6$$

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

$$K = \begin{bmatrix} f_x & \theta & o_x \\ 0 & f_y & o_y \\ 0 & 0 & 1 \end{bmatrix}$$





TILT – Camera Calibration with Radial Distortion

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

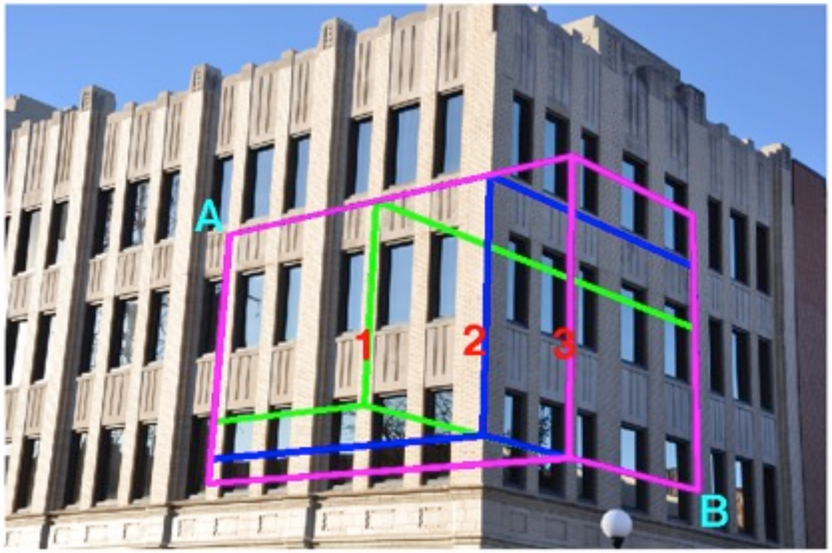
$$\tau_0 = (K, K_c), \quad \tau_i = (R_i, T_i).$$

Previous approach

Low-rank method

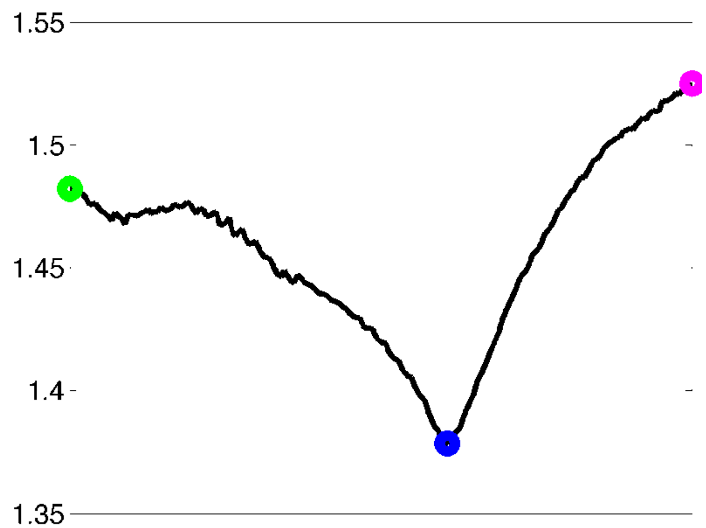
Zhang, Matsushita, and Ma, in CVPR 2011

TILT – Holistic 3D Reconstruction of Urban Scenes



$$\min \|A\|_* + \|E\|_1 \quad \text{s.t.}$$

$$A + E = [D_1 \circ \tau_1, D_2 \circ \tau_2]$$

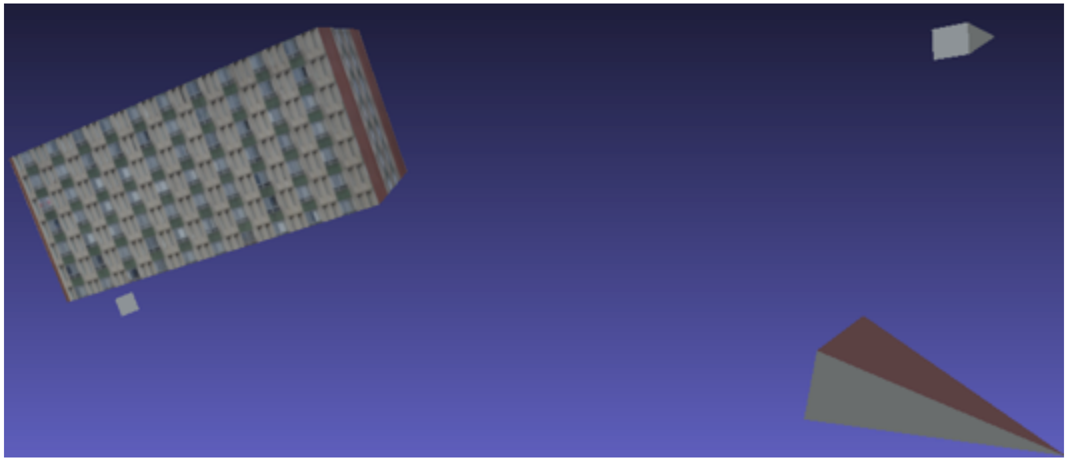


Mobahi, Zhou, and Ma, in ICCV 2011

TILT – Holistic 3D Reconstruction of Urban Scenes

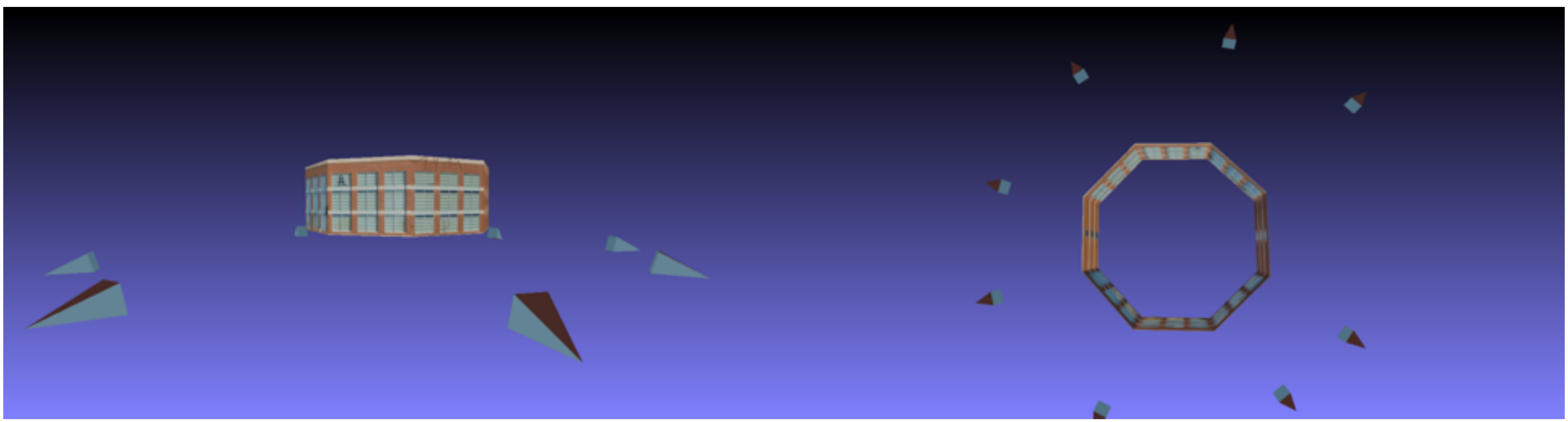
From one input image

From four input images



TILT – Holistic 3D Reconstruction of Urban Scenes

From eight input images



Virtual reality in urban scenes

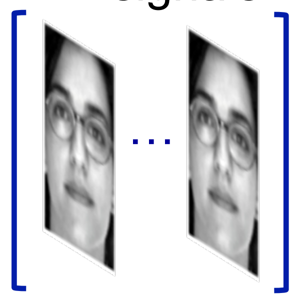
Registering Multiple Images – Robust Alignment

D – corrupted & misaligned observation

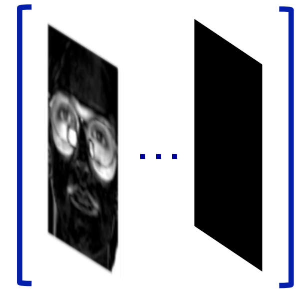
A – aligned low-rank signals

E – sparse errors

$\circ\, \tau \;=$



$+$



Problem: Given $D \circ \tau = A_0 + E_0$, recover τ , A_0 and E_0 .

Parametric deformations (rigid, affine, projective…)

Low-rank component

Sparse component

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)

Average

RASL – Faces Aligned

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

Average

RASL – Faces Repaired and Cleaned

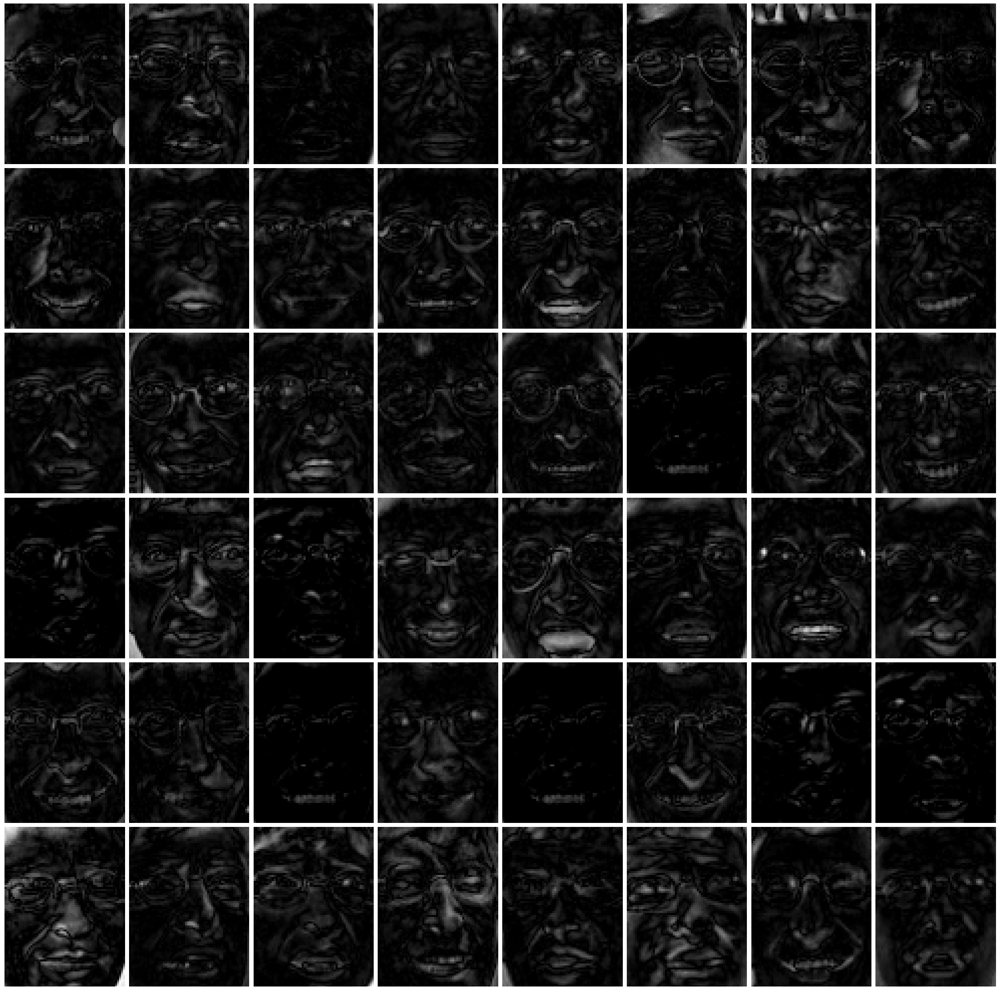
Output: clean low-rank faces (A)

Average

Peng, Ganesh, Wright, Ma, CVPR'10, TPAMI'11

RASL – Sparse Errors of the Face Images

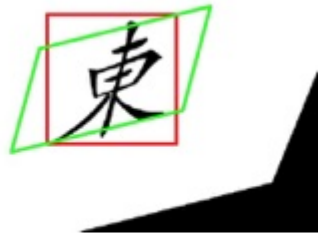
Output: sparse error images (E)

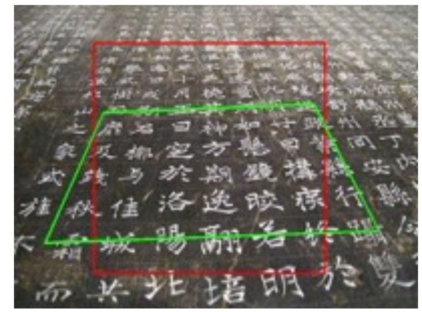


Peng, Ganesh, Wright, Ma, CVPR'10, TPAMI'11

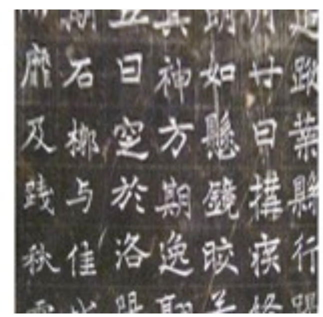
Object Recognition – Regularity of Texts at All Scales!

Input (red window D)





Output (rectified green window A)



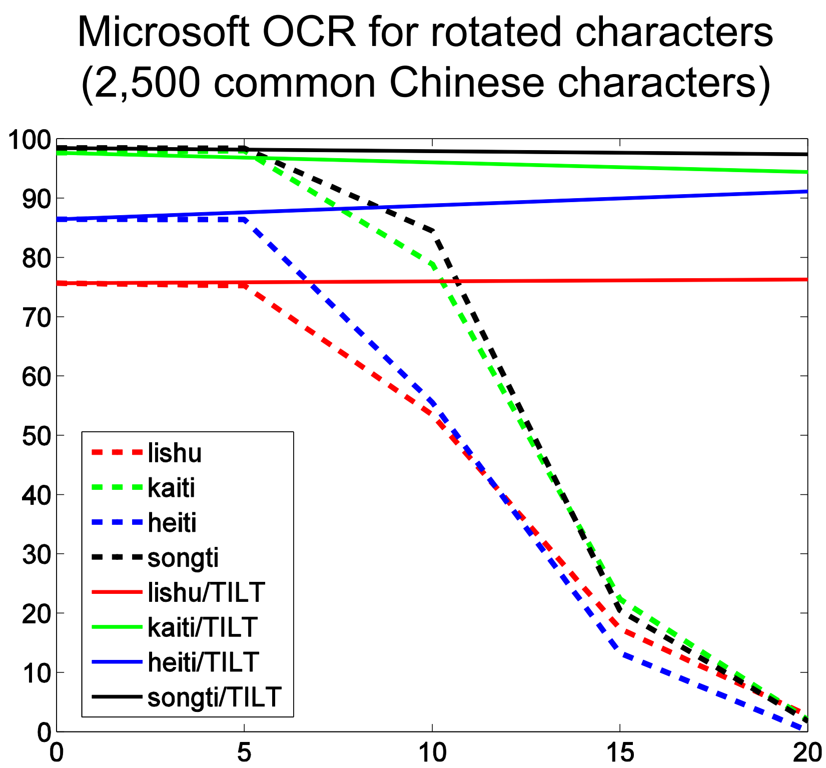
Recognition – Street Sign Rectification

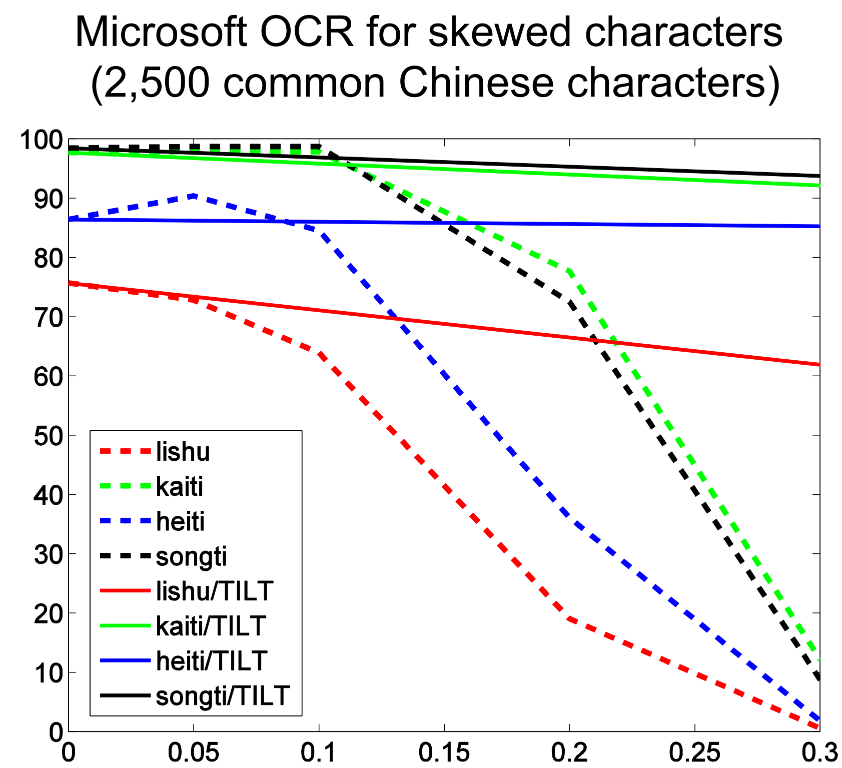
$A_1 \quad A_2 \quad A_3 \quad A_4$

$$\min \sum_{i=1}^{4} \|A_i\|_* + \lambda \|E_i\|_1$$

$$\text{subj} \;\; D \circ \tau = [A_1 \cdots A_4] + [E_1 \cdots E_4].$$

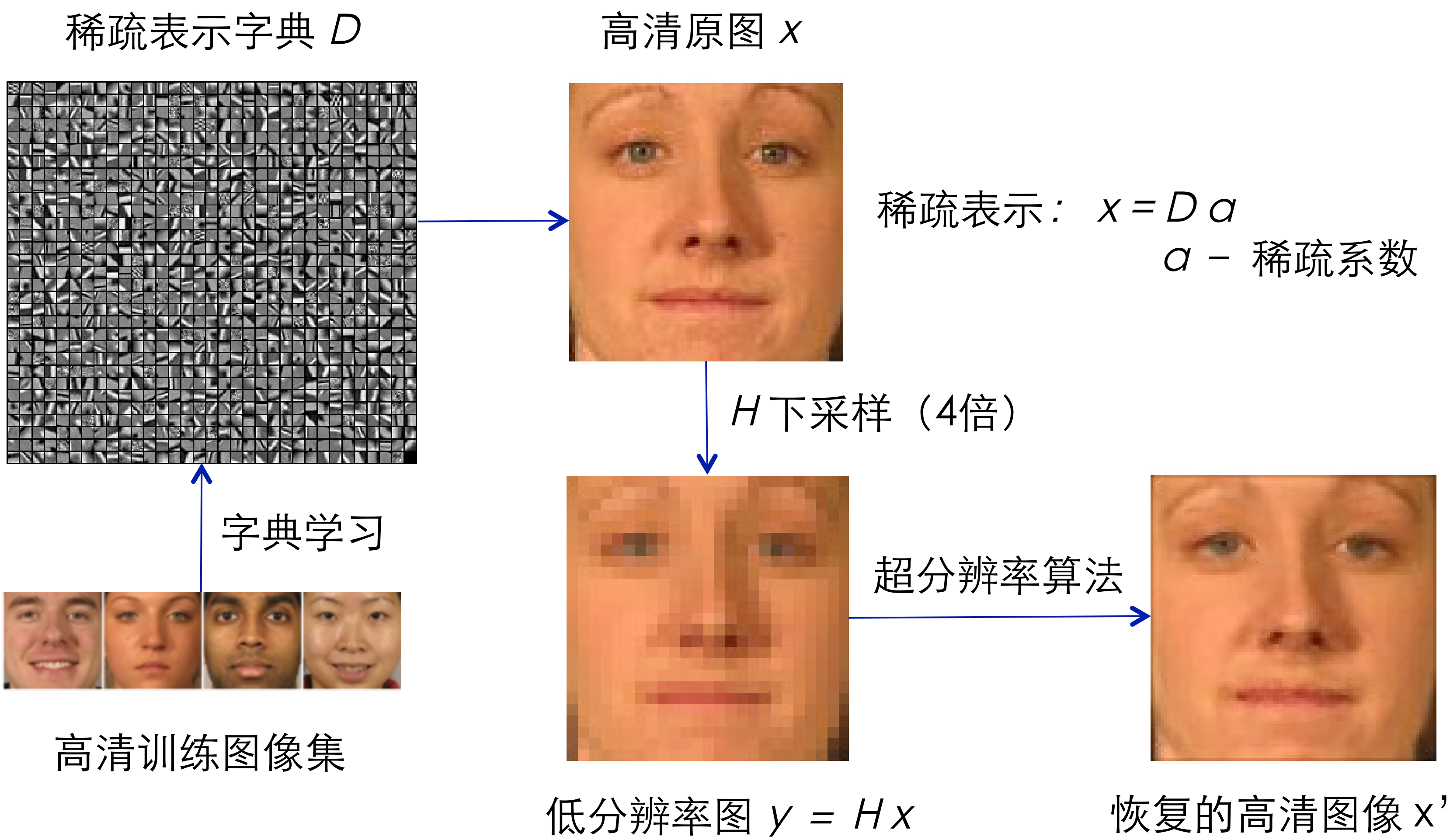
Recognition – Character Rectification and Recognition





Xin Zhang, Zhouchen Lin, and Ma, ICDAR 2013

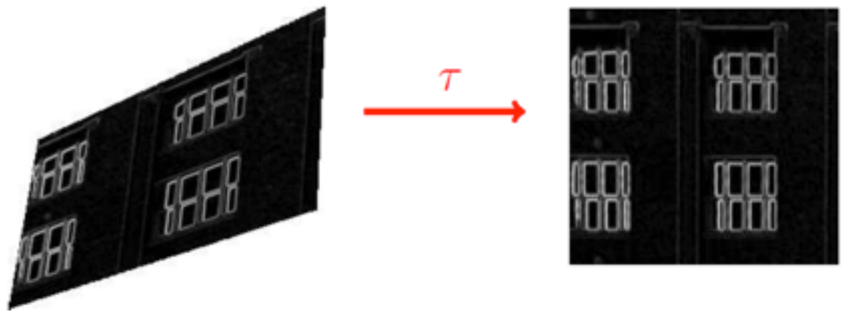
Super Resolution via Transform Invariant Group Sparsity



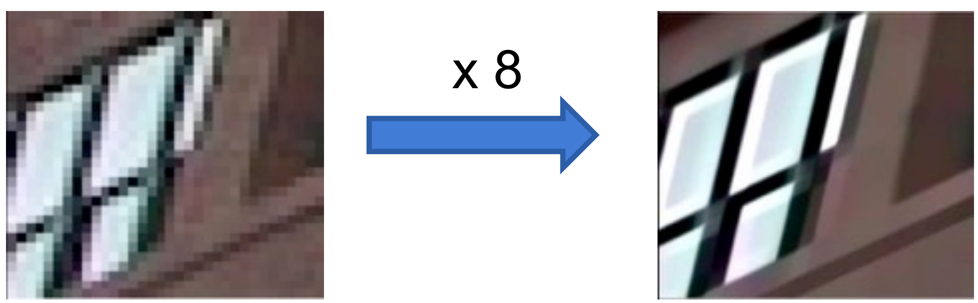
Super Resolution via Transform Invariant Group Sparsity

Aim: Exploiting non-local structures to perform super-resolution at large upsampling factors by

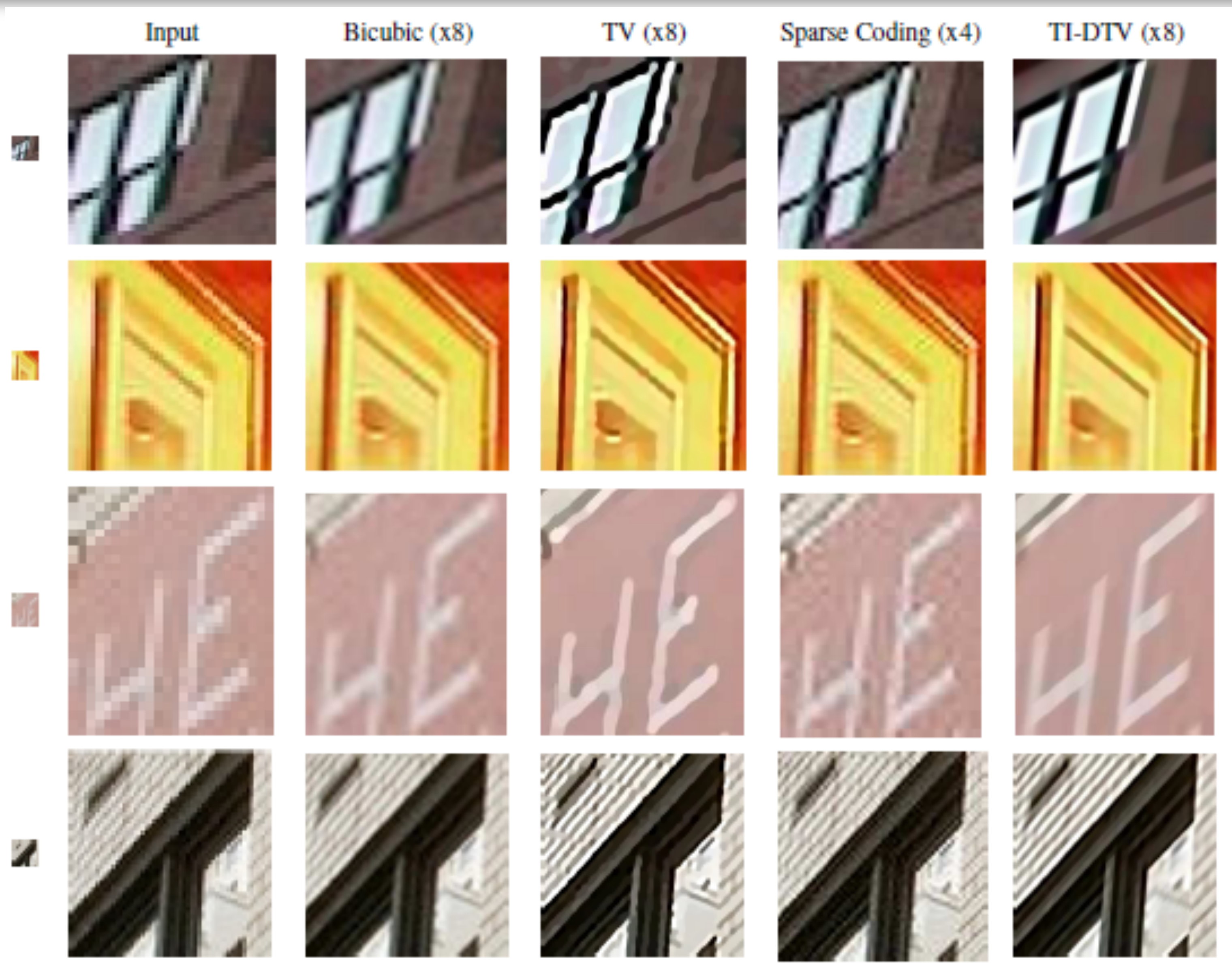
1. Learning the transformation that reveals the group-sparse structure of the image gradient (via TILT)



2. Enforcing this structure through group-sparse regularizers (DTV) that incorporates the transform and is consequently invariant to the transform



Super Resolution via Transform Invariant Group Sparsity



Carlos Fernandez and Emmanuel Candes of Stanford, ICCV2013

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;

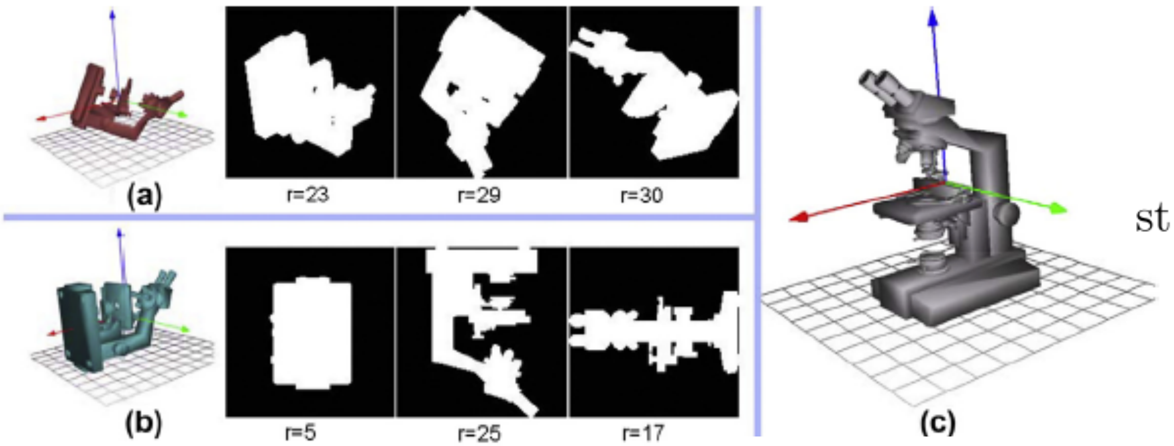
4. These algorithms start to exert significant impact on **image/video processing, 3D reconstruction, and object recognition**.

… …

But try not to abuse or misuse them…

Other Applications – Upright orientation of man-made objects

TILT for 3D: Unsupervised upright orientation of man-made 3D objects



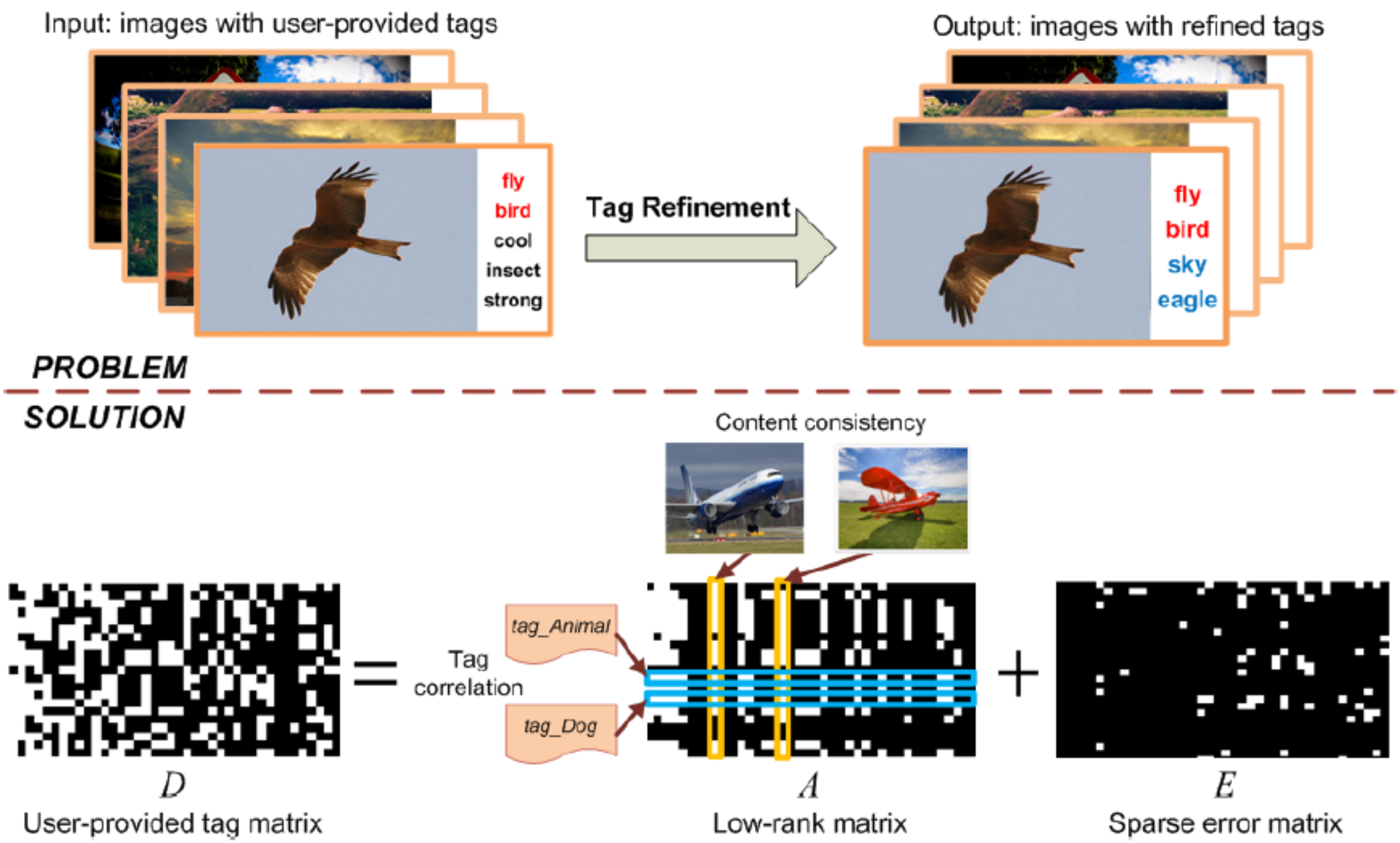
$$\min \sum_{i=1}^{3} \|A_i\|_* + \lambda \|E_i\|_1$$

$$\text{st } D \circ \tau = [A_1, A_2, A_3] + [E_1, E_2, E_3].$$

Fig. 10. More models which have been successfully tested through our algorithm.

Jin, Wu, and Liu of USTC, China, Graphical Models, 2012.

Other Data/Applications – Web Image/Tag Refinement



Other Data/Applications – Web Document Corpus Analysis

Latent Semantic Indexing: the classical solution (PCA)

Documents

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 "re° ect not only our out-standing performance over the past few years but also our optimism about the company's future."

D

d_{ij} word frequency (or TF/IDF)

$$D = A + Z$$
$$= U_1 \Sigma_1 V_1^T + \underline{U_2 \Sigma_2 V_2^T}$$

Dense, difficult to interpret

a better model/solution?

$$D = A + \underline{E}$$

Low-rank
"background"
topic model

Informative,
discriminative
"keywords"

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 "re°ect not only our outstanding performance over the past few years but also our optimism about the company's future."

Other Data/Applications – Protein-Gene Correlation

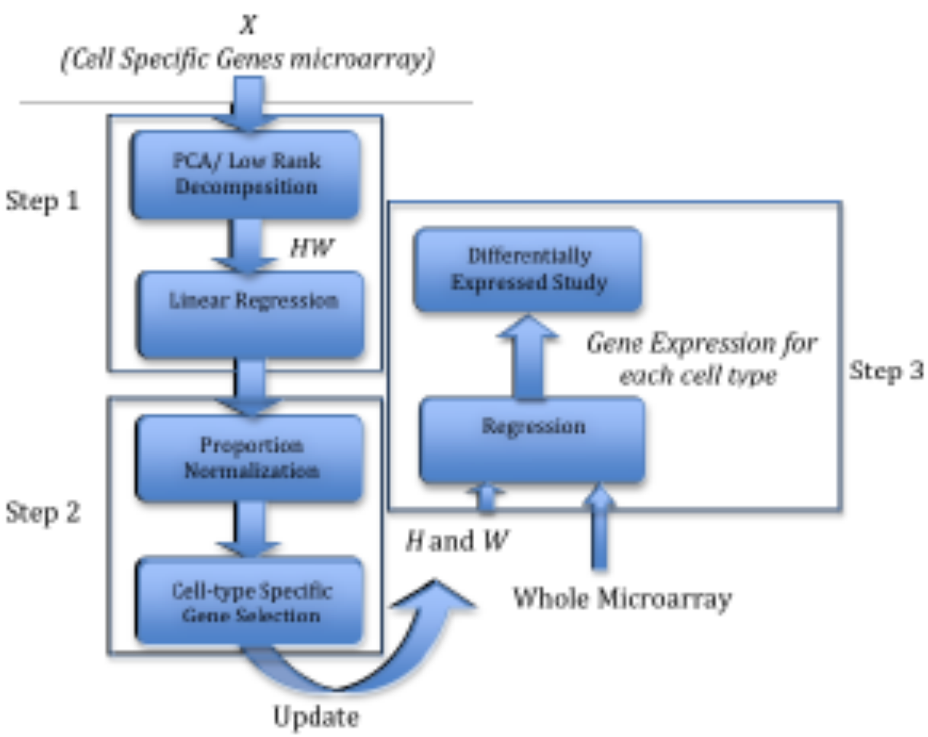


Fig. 1. The diagram of the workflow of the method presented in this paper.

Microarray data

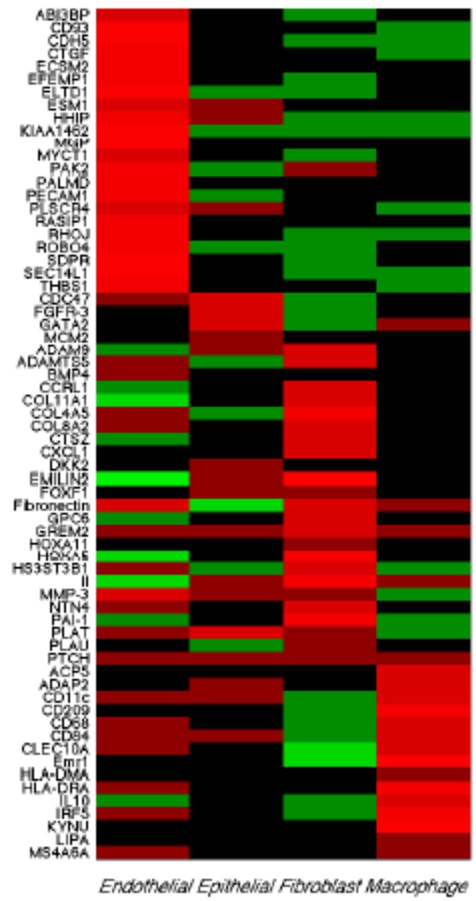
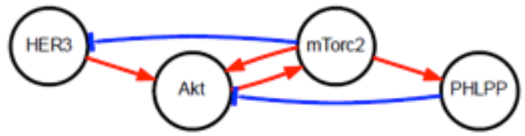


Fig. 6. HeatMap of estimated gene signatures for the sorted cell specific genes after adjustments based on fold changes. RPCA is used in the first step. It is clear that this matrix is close to a block diagonal structure.

Wang, Machiraju, and Huang of Ohio State Univ. , Bioinformatics.

Other Data – Time Series Gene Expressions



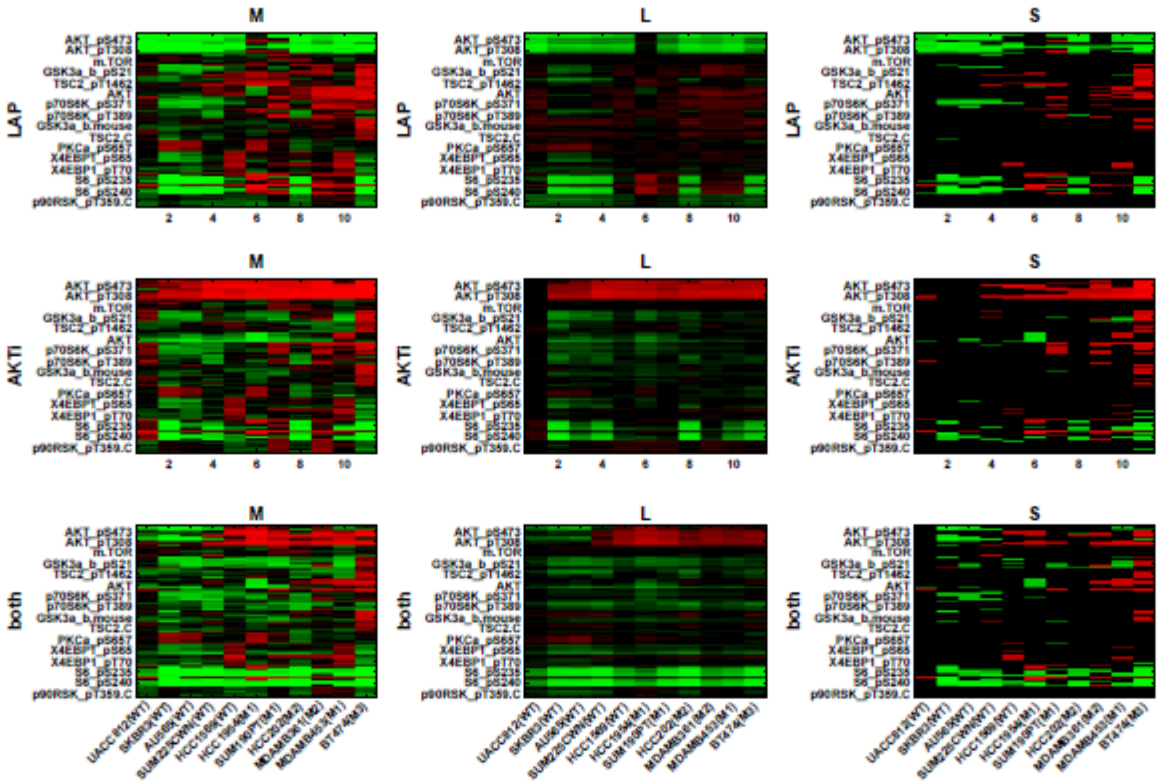
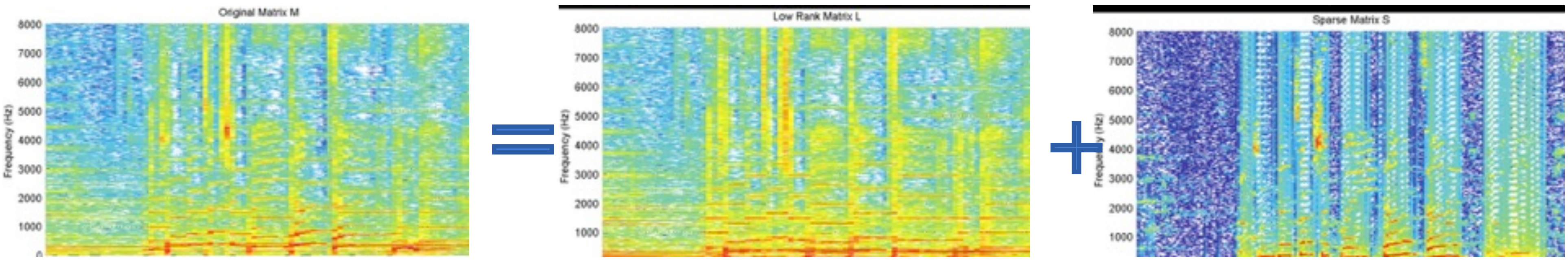
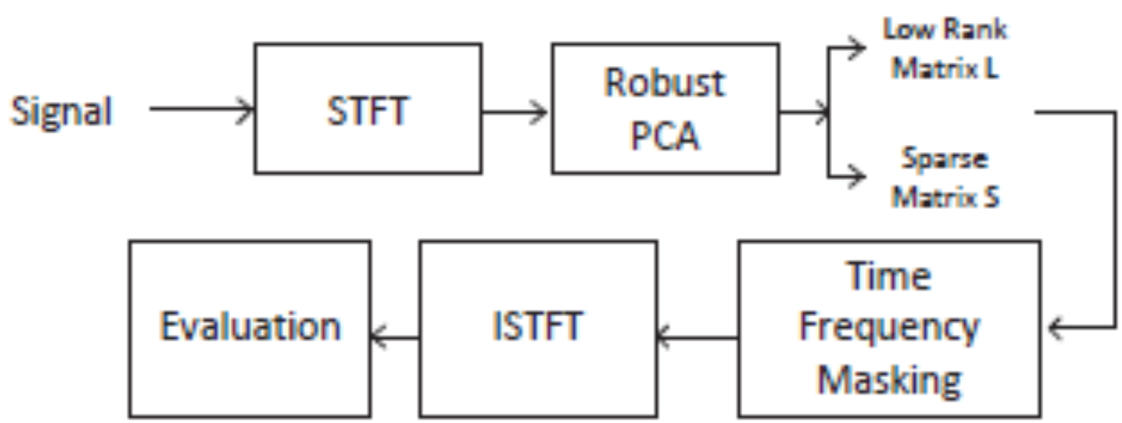


Figure S4. Separation result: (1_{st} column) raw data (2_{nd} column) low-rank component and (3_{rd} column) highly corrupted sparse component using threshold (M1: H1047R (kinase domain mutation) M2: E545K (helical domain mutation), and M3: K111N mutation in PIK3CA).

Other Data/Applications – Lyrics and Music Separation

Songs (STFT) | Low-rank (music) | Sparse (voices)





Po-Sen Huang, Scott Chen, Paris Smaragdis, Mark Hasegawa-Johnson of UIUC, ICASSP 2012.

Other Data/Applications – Internet Traffic Anomalies

Network Traffic = Normal Traffic + Sparse Anomalies + Noise

$$D = L + RS + N$$

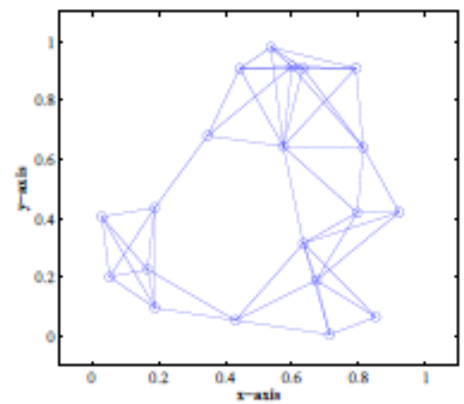
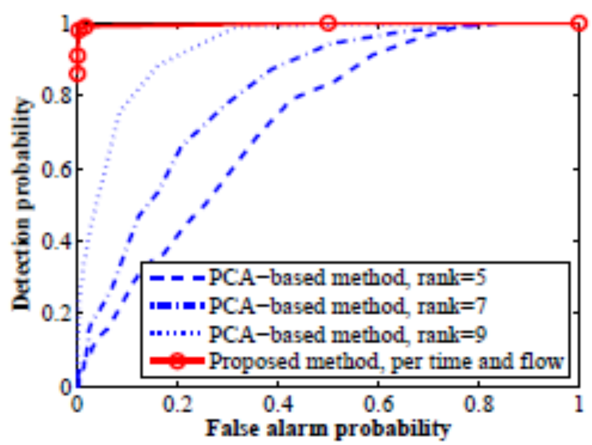
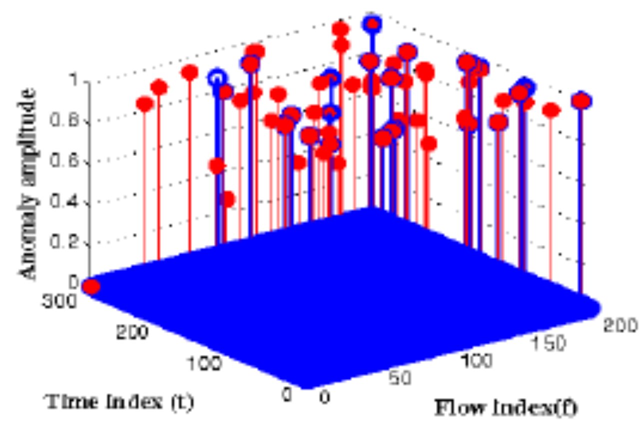


Fig. 2. Network topology graph.





Mardani, Mateos, and Giannadis of Minnesota, Trans. Information Theory, 2013.

Other Data/Applications – Robust Filtering and System ID

GPS on a Car:

$$\begin{cases} \dot{x} &= Ax + Bu, \quad A \in \Re^{r\times r} \\ y &= Cx + z + {\color{red}e} \end{cases}$$

gross sparse errors
(due to buildings, trees…)

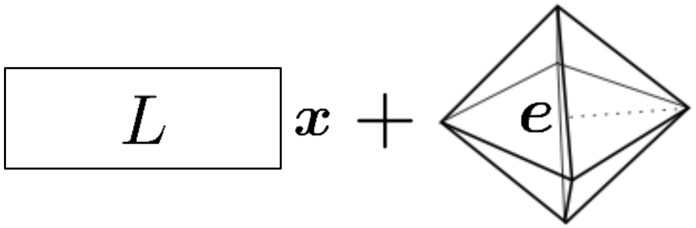
Robust Kalman Filter: $\hat{x}_{t+1} = Ax_t + K(y_t - C\hat{x}_t)$

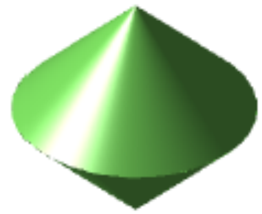
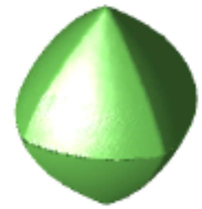
Robust System ID: $$\underbrace{\begin{bmatrix} y_n & y_{n-1} & y_{n-2} & \cdots & y_0 \\ y_{n-1} & y_{n-2} & \cdots & \ddots & y_{-1} \\ y_{n-2} & \cdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & y_{-n+2} \\ y_0 & y_{-1} & \cdots & y_{-n+2} & y_{-n+1} \end{bmatrix}}_{\text{Hankel matrix}} = \mathcal{O}_{n\times r} X_{r\times n} + S$$

CONCLUSIONS – A Unified Theory for Sparsity and Low-Rank

	Sparse Vector	***Low-Rank Matrix***
Low-dimensionality of	individual 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	$Y = A(X) + B(E) + Z$	

Compressive Sensing of Low-Dimensional Structures




A norm $\|\cdot\|$ is said to be **decomposable** at $\boldsymbol{X}$ if there exists a subspace T and a matrix $\boldsymbol{S}$ such that

$$\partial\|\cdot\|(\boldsymbol{X}) = \{\Lambda \mid \mathcal{P}_T(\Lambda) = \boldsymbol{S}, \|P_{T^\perp}(\Lambda)\|^* \le 1\},$$

where $\|\cdot\|^*$ is the dual norm of $\|\cdot\|$, and $\mathcal{P}_{T^\perp}$ is nonexpansive w.r.t. $\|\cdot\|^*$.

Theorem [Candes, Recht'11] Any low-complexity signal $\boldsymbol{X}^0$ can be exactly recovered from high compressive measurements via convex optimization:

$$||\boldsymbol{X}||_\diamond \quad \text{subject to} \quad \mathcal{P}_Q(\boldsymbol{X}) = \mathcal{P}_Q(\boldsymbol{X}^0),$$

for a decomposable norm $\|\cdot\|_\diamond$.

Compressive Sensing and Unmixing of Low-dim Structures

Suppose $(\boldsymbol{X}_1^0, \ldots, \boldsymbol{X}_k^0) \;=\; \arg\min \sum_{i=1}^k \lambda_i \|\boldsymbol{X}_i\|_{(i)} \quad \text{subj} \quad \sum_{i=1}^k \boldsymbol{X}_i = \sum_{i=1}^k \boldsymbol{X}_i^0,$
for decomposable norms $\|\cdot\|_{(i)}$ that majorize the Frobenius norm.

Theorem 6 (Compressive Sensing of Mixed Low-Comp. Structures). *Let $Q^\perp$ be a random subspac of $\mathbb{R}^{m\times n}$ of dimension*

$$\dim(Q) \geq O(\log^2 m) \times \text{intrinsic degrees of freedomof } (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k),$$

distributed according to the Haar measure, independent of $\boldsymbol{X}_i$. Then with very high probability

$$(\boldsymbol{X}_1^0, \ldots, \boldsymbol{X}_k^0) \;=\; \arg\min \sum_{i=1}^k \lambda_i \|\boldsymbol{X}_i\|_{(i)} \quad \text{subj} \quad \mathcal{P}_Q\Big[\sum_{i=1}^k \boldsymbol{X}_i\Big] = \mathcal{P}_Q\Big[\sum_{i=1}^k \boldsymbol{X}_i^0\Big],$$

and the minimizer is unique.

Extensions – 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 '11'12] – 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,Amenlunxen+McCoy+Tropp'13] – phase transition for recovery and decomposition of structures

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

- [Wright+Ganesh+Min+Ma, ISIT'12,I&I'13] – compressive superposition of decomposable structures

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

Take home message: ***Let the data and application tell you the structure…***

Relationships with Deep Neural Networks

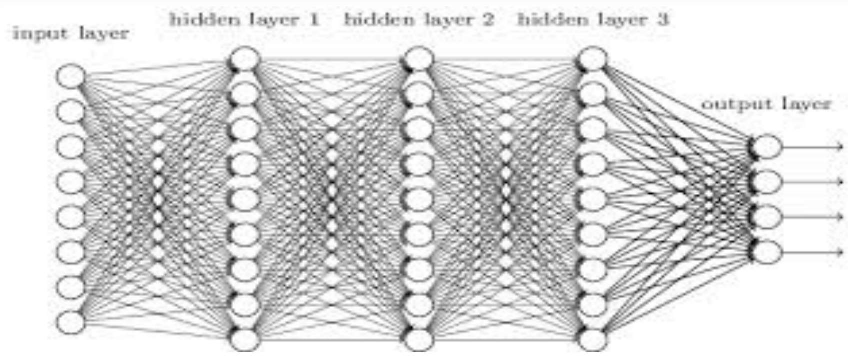
1. **Evolution of the Structures of Deep Networks**
   FNN -> CNN -> ResNet -> ???

2. **Deep Learning and Sparsity**
   Cascaded Structured Matrix Factorization
   Global Optimality of Training

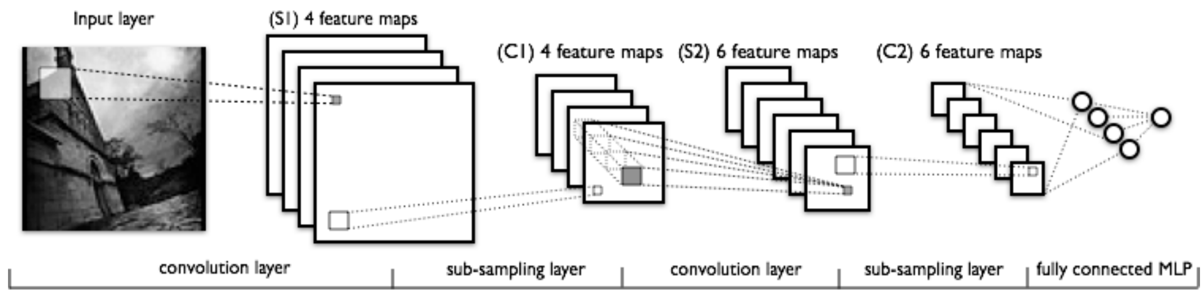
3. **Supervision versus No-supervision**
   Simple Shallow Networks by Design
   PCANet (and ScatteringNet and CapsuleNet)

Evolution of DNN – More Principled Structures

Fully Connected
Neural Networks
(before 2011)



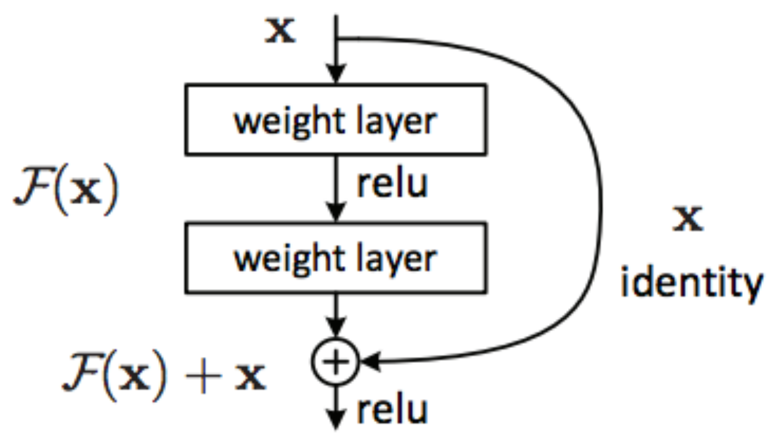
Convolution
Neural Networks
(AlexNet 2011)



translational invariance

Residual
Neural Networks
(2015, and AlphaGo Zero)

constrained optimization?



Temporally coherent Sparse Coding for Anomaly Detection in Video

$$\min_{A,\alpha_t} \sum_{t=1}^{T} \|x_t - A\alpha_t\|_2^2 + \lambda_1\|\alpha_t\|_1 + \lambda_2 S_{t,t-1}\|\alpha_t - \alpha_{t-1}\|_2^2$$

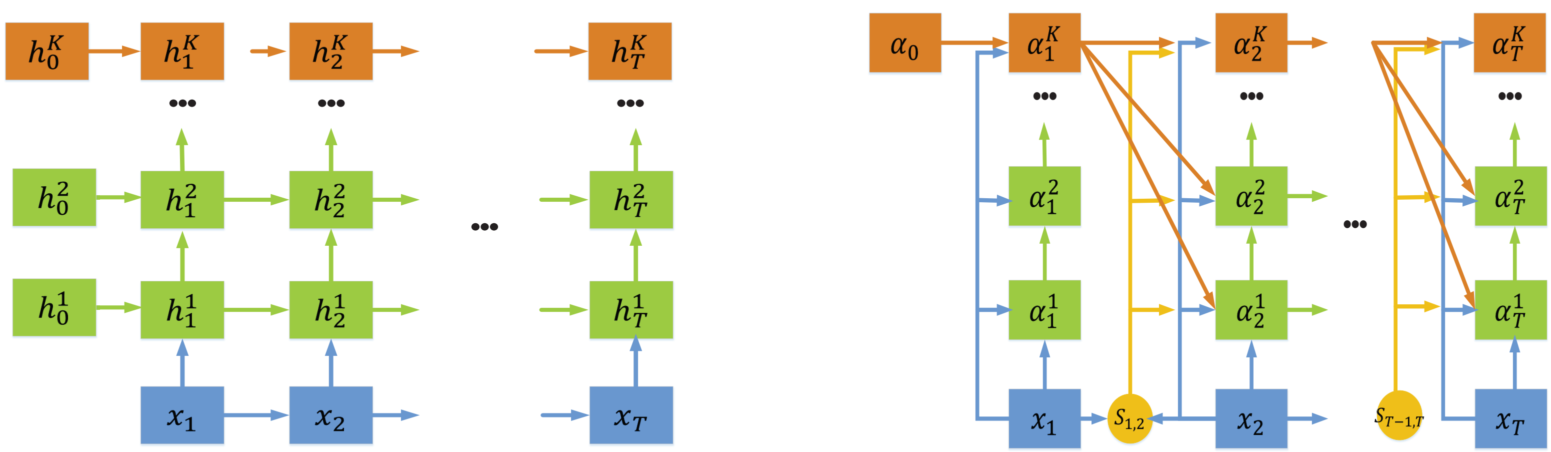
$$\text{s.t.} \quad \|A(:,i)\| \leq 1$$

Algorithm 1 Sequential iterative soft-thresholding algorithm.

Input: extracted feature $x_{1:T}$, hyper-parameter $\lambda_1, \lambda_2, \gamma$, initial $\hat{\alpha}_0$, the steps of ISTA K

```
for t = 1 to T do
    α̂_t^0 = α_{t-1}
    for k = 1 to K do
        z = [I − (1/γ)(A^T A + S_{t-1,t} λ_2 I)] α̂_t^{k-1} + (1/γ) A^T x_t
        α̂_t^{(k)} = soft_{λ_1/γ}(z + (S_{t-1,t} λ_2 / γ) α_{t-1})
    end for
    α_t = α̂_t^K
end for
return α_{1:T};
```

RNN structures derived from group sparsity!



(a) Vanilla stacked RNN [26]

(b) Stacked RNN couterpart of TSC

Figure 1. The blue boxes represent the input x_t of stacked RNNs. The green and orange boxes represent coding vectors α_t^k. The yellow circles are similarities between neighboring frames.

Structured Attentions for Visual Question Answering

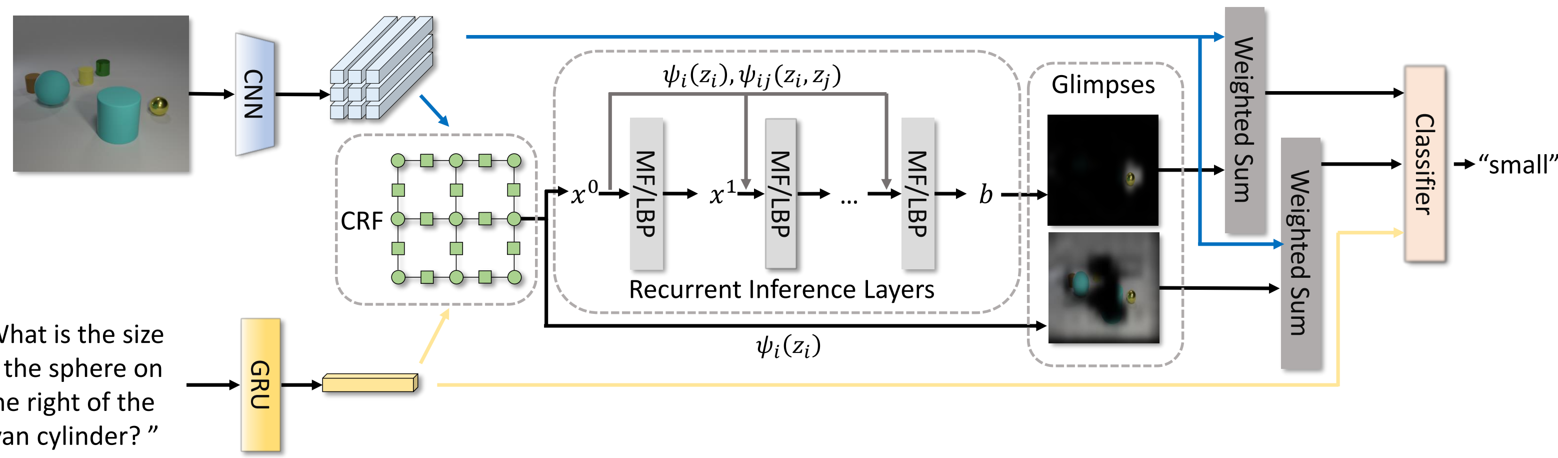
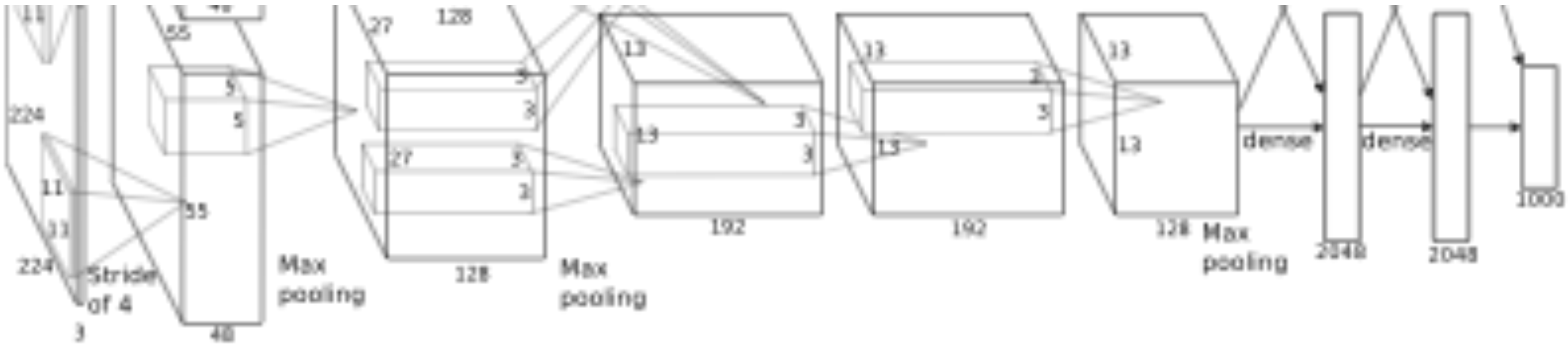


Figure 2. The whole picture of the proposed model. The unary potential $\psi_i(z_i)$ and pairwise potential $\psi_{ij}(z_i, z_j)$ are computed with Eq. (8), which are inputs to the recurrent inference layers. $\psi_i(z_i)$ is also used as an additional glimpse, which usually detects the key-word objects. In the inference layers, x^i represents $b^{(i)}$ for MF and $m^{(i)}$ for LBP. The recurrent inference layers generates a refined glimpse with Mean Field or Loopy Belief Propagation. The 2 glimpses are used to weight-sum the visual feature vectors. The classifier use both of the attended visual features and the question feature to predict the answer. The demonstrated image is a real case.

Recurrent inference layers derived from MF or LBP (for graphical models)!

II. Deep Learning and Sparsity

- Deep learning is a cascaded matrix factorization



$$\Phi(X^1,\ldots,X^K) = \psi_K(\cdots\psi_2(\psi_1(VX^1)X^2)\cdots X^K)$$

nonlinearity features weights

$$\min_{X^1,\ldots,X^K} \ell(Y,\Phi(X^1,\ldots,X^K)) + \lambda\Theta(X^1,\ldots,X^K)$$

loss labels regularizer

Deep Learning and Sparsity

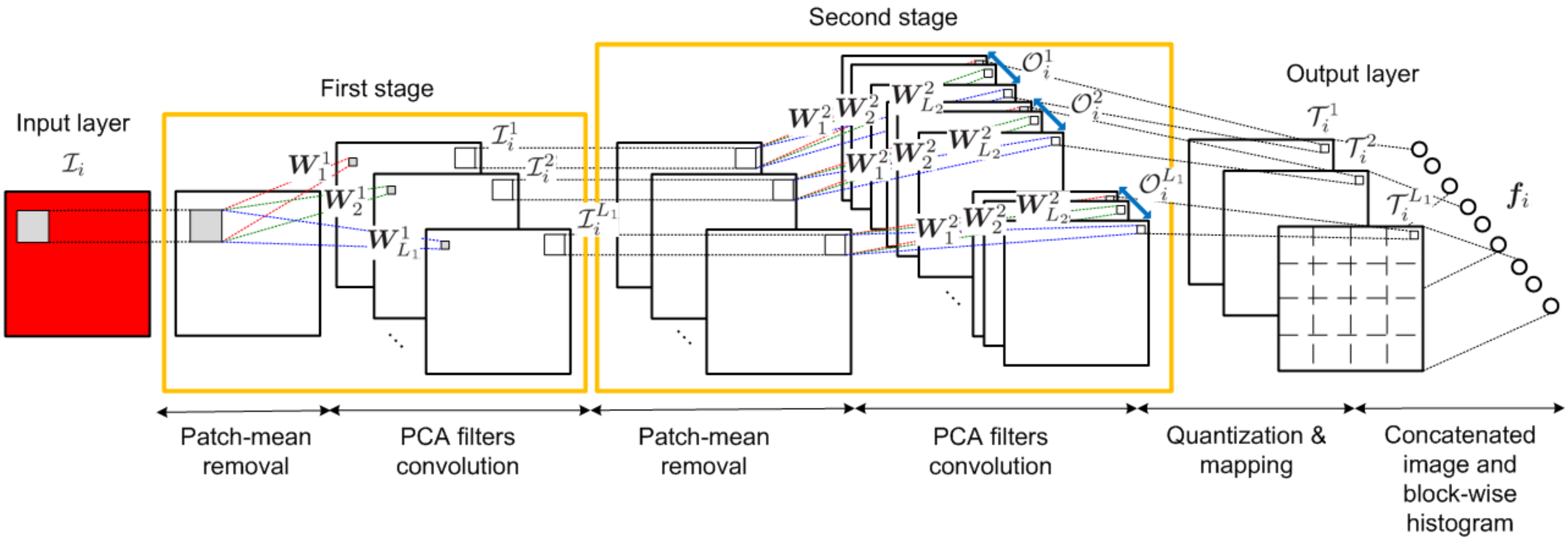
$$\min_{X^1,\ldots,X^K} \ell(Y, \Phi(X^1,\ldots,X^K)) + \lambda\Theta(X^1,\ldots,X^K)$$

- **Theorem:** If the functions Φ and Θ are sums of positively homogeneous functions, then any local minimizer such that for some *i* and all k $X_i^k = 0$ gives a global minimizer

- Examples of positively homogeneous compositions Φ
  - Matrix multiplication: matrix factorization
  - CANDECOMP/PARAFAC decompositions: tensor factorization
  - Rectified linear units + max pooling: deep learning

- Examples of positively homogeneous regularizers Θ
  - Sums of products of norms (L1, L2, TV, etc.): structured factorizations

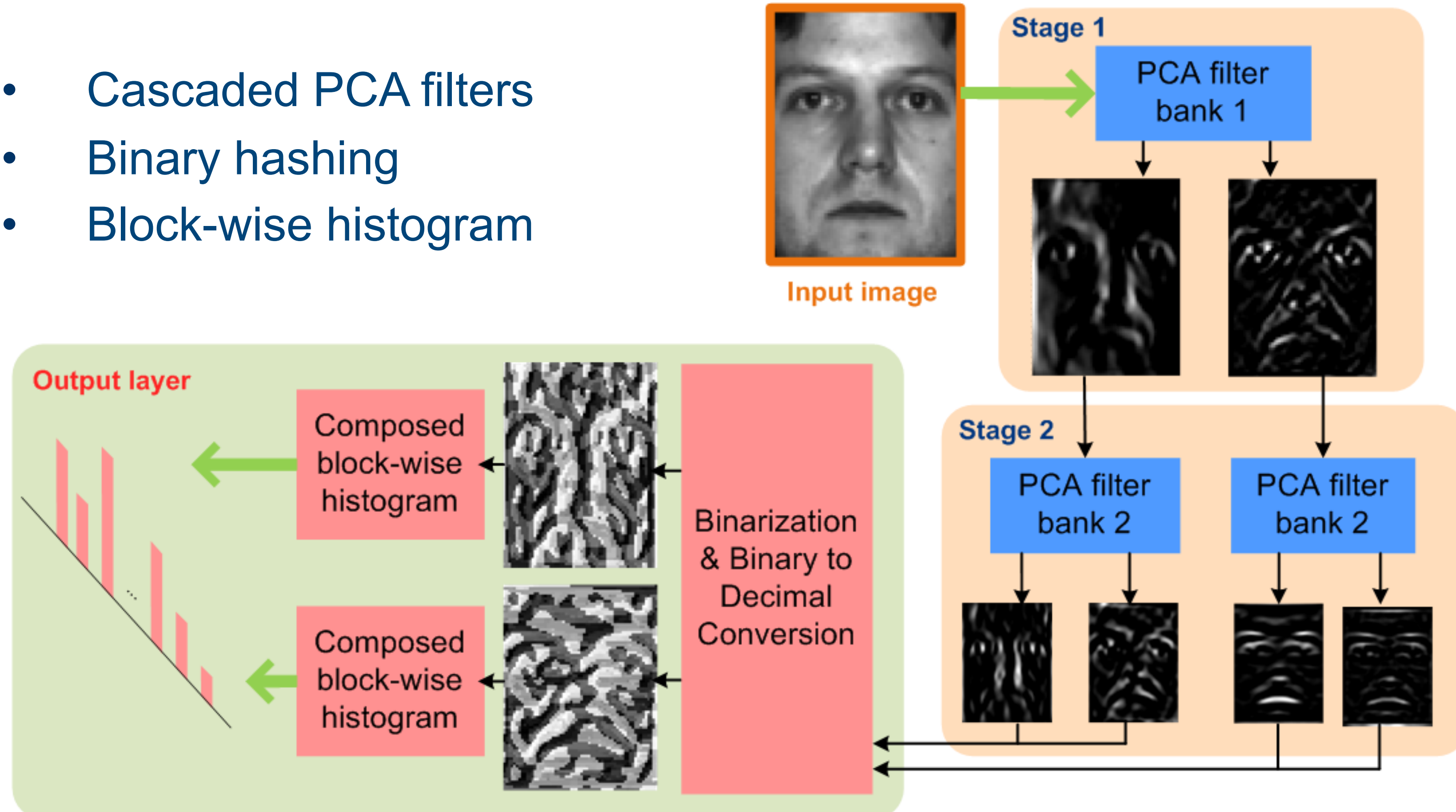
III. Supervision or None? - PCANet



2-3 layers, fixed topology, simplest data-adaptive linear mapping, and simplest nonlinear processing and simplest pooling…

PCANet – Basic Structure

- Cascaded PCA filters
- Binary hashing
- Block-wise histogram



- **ScatteringNet (S. Mallat et. al. 2013)**
- **CapsuleNet (G. Hinton et. al. 2017)**

- Two to three layers!
- By design, no supervision!
- Pure feed-forward, no BP!

PCANet, *Chan and Ma et. al. in IEEE Trans. On Image Processing 2015*

PCANet – Test on NIST FERET

FERET contains images of 1,196 different individuals with up to 5 images of each individual.

The probe set is divided into four subsets

- *Fb* with different expression changes;
- *Fc* with different lighting conditions;
- *Dup-I* taken within the period of three to four months;
- *Dup-II* taken at least one and a half year apart.

Gallery Fb Fc

Gallery Dup-I Dup-II

PCANet, Chan and Ma et. al. in IEEE Trans. On Image Processing 2015

PCANet – Test on NIST FERET

The non-overlapping block size (for histogram) is 15x15.

The dimension of the PCANet features are reduced to 1000 by a whitening PCA (WPCA).

“Trn. CD” means trained with standard FERET CD dataset

The NN classifier with cosine distance is used.

Recognition rates (%) on FERET dataset.

Probe sets	*Fb*	*Fc*	*Dup-I*	*Dup-II*	Avg.
LBP [18]	93.00	51.00	61.00	50.00	63.75
DMMA [25]	98.10	98.50	81.60	83.20	89.60
P-LBP [21]	98.00	98.00	90.00	85.00	92.75
POEM [26]	99.60	99.50	88.80	85.00	93.20
G-LQP [27]	**99.90**	**100**	93.20	91.00	96.03
LGBP-LGXP [28]	99.00	99.00	94.00	93.00	96.25
sPOEM+POD [29]	99.70	**100**	94.90	94.00	97.15
GOM [30]	**99.90**	**100**	95.70	93.10	97.18
PCANet-1 (Trn. CD)	99.33	99.48	88.92	84.19	92.98
PCANet-2 (Trn. CD)	99.67	99.48	**95.84**	**94.02**	**97.25**
PCANet-1	99.50	98.97	89.89	86.75	93.78
PCANet-2	99.58	**100**	95.43	**94.02**	**97.26**

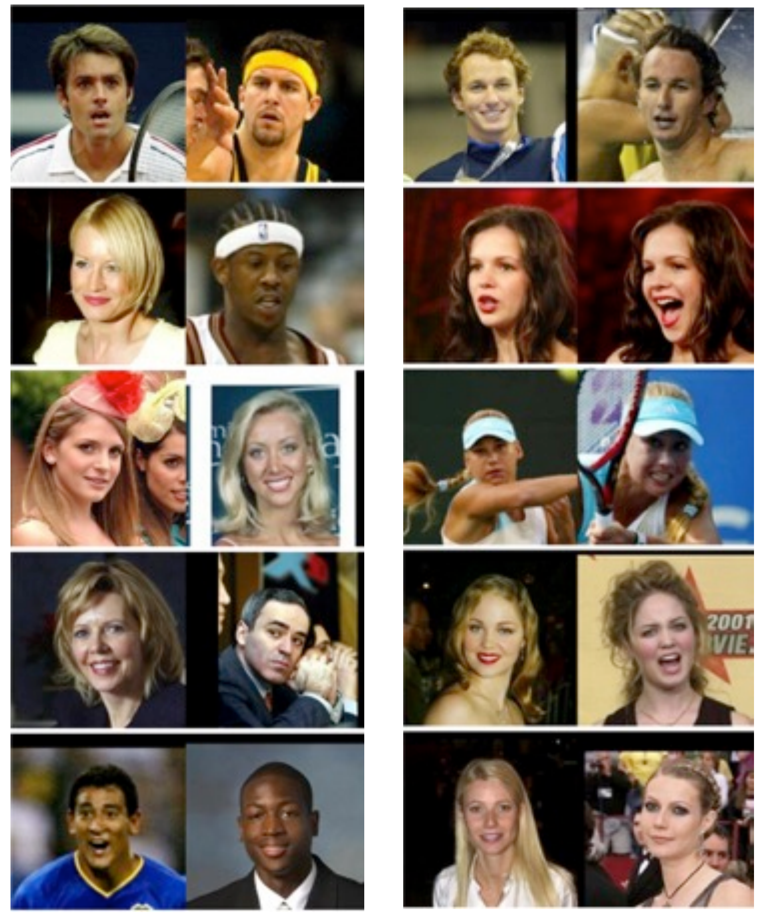
PCANet – Test on LFW

LFW contains 13,233 face images of 5,749 individuals, collected from the web.

We use LFW-a [aligned version].

“Unsupervised” setting.

- View 1 dataset is used to learn the PCA filters and the projection matrix of the WPCA, and to decide a matching threshold.
- the trained PCANet is applied to View 2 dataset, 10 subsets of pairs.



Mismatched pairs **Matched pairs**

PCANet – Test on LFW

PCANet parameters: the filter size k1 = k2 = 7, the number of filters L1 = L2 = 8, and block size is 15x13.

The features of PCANet-1 and PCANet-2 are projected onto 400 and 3,200 dimensions, respectively.

"sqrt" means PCA features followed with a square-root operation.

We use NN classifier with cosine distance.

Comparison of verification rates (%) on LFW under unsupervised setting.

Methods	Accuracy
POEM [26]	82.70±0.59
High-dim. LBP [36]	84.08
High-dim. LE [36]	84.58
SFRD [37]	84.81
I-LQP [27]	86.20±0.46
OCLBP [33]	86.66±0.30
PCANet-1	81.18 ± 1.99
PCANet-1 (sqrt)	82.55 ± 1.48
PCANet-2	85.20 ± 1.46
PCANet-2 (sqrt)	86.28 ± 1.14

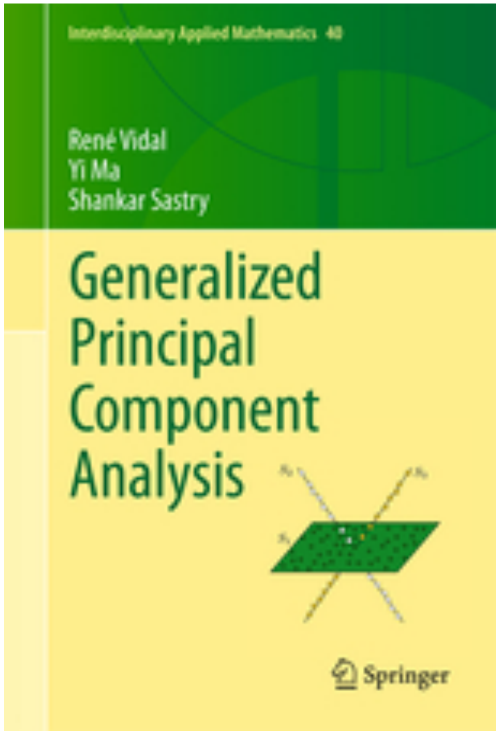
REFERENCES

From Supervised to Unsupervised Learning,
From One Subspace to Multiple Subspaces.

A recent book:

- ***Generalized Principal Component Analysis***

  R. Vidal, Yi Ma, S. Sastry, Springer 2016



REFERENCES

Core References:

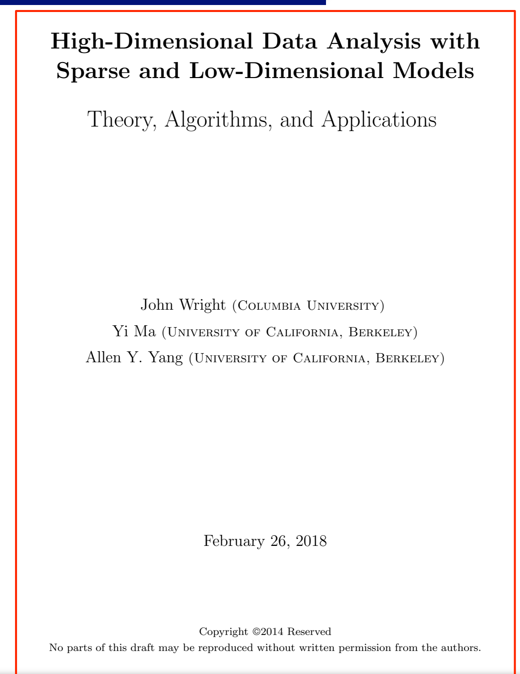
- *Robust Principal Component Analysis*? Candes, Li, Ma, Wright, Journal of the ACM, 2011.
- *TILT: Transform Invariant Low-rank Textures,* Zhang, Liang, Ganesh, and Ma, IJCV 2012.
- *Compressive Principal Component Pursuit*, Wright, Ganesh, Min, and Ma, IMA I&I 2013.

Website (codes, applications, & references):

http://perception.csl.illinois.edu/matrix-rank/home.html

A New Graduate Textbook:

High-Dimensional Data Analysis with Sparse and Low-Dimensional Models
-- Theory, Algorithms, Applications


High-Dimensional Data Analysis with Sparse and Low-Dimensional Models

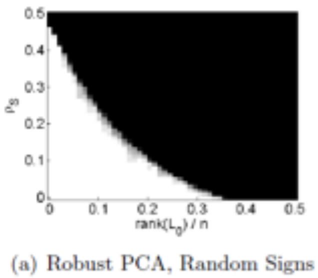
Theory, Algorithms, and Applications

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

February 26, 2018

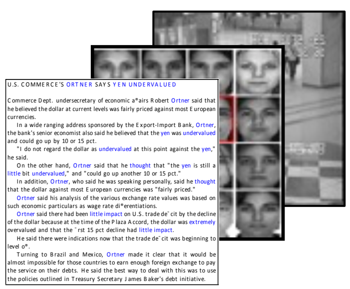
Copyright ©2014 Reserved
No parts of this draft may be reproduced without written permission from the authors.

A Perfect Storm…



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

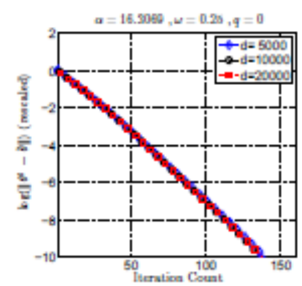
BIG DATA
(images, videos, voices, texts, biomedical, geospatial, consumer data…)



Cloud Computing
(parallel, distributed, scalable platforms)

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

Computational Methods
(convex optimization, first-order algorithms, random sampling, deep networks…)



ACKNOWLEDGEMENT

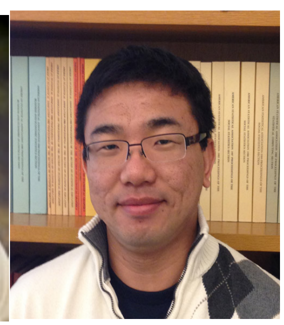
Colleagues:

- Prof. Emmanuel Candes (Stanford)
- Prof. Zhouchen Lin (MSRA, now Peking Univ.)
- Prof. Yasuyuki Matsushita (MSRA, now Osaka Univ.)
- Prof. Shuicheng Yan (Na. Univ. Singapore)
- Prof. Lei Zhang (HK Polytech Univ.)
- Prof. Liangshen Zhuang (USTC)
- Prof. Weisheng Dong (Xidian Univ., China)
- Prof. Shenghua Gao (ShanghaiTech Univ., China)
- Prof. Rene Vidal (Johns Hopkins Univ.)

Students:

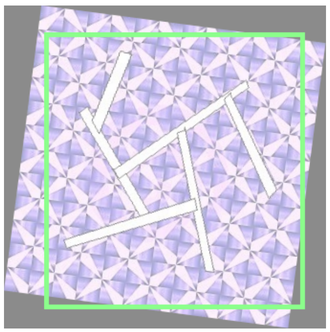
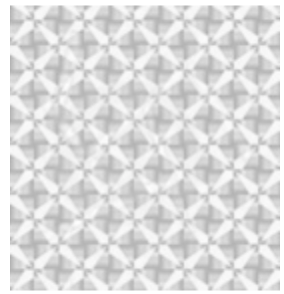
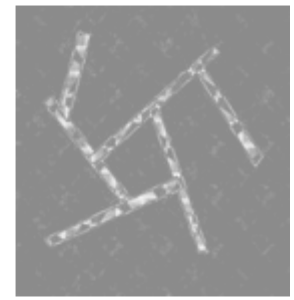
- John Wright (UIUC, now Columbia)
- Arvind Ganesh (UIUC, now Google)
- Zhengdong Zhang (MSRA, now MIT)
- Xiao Liang (MSRA, Tsinghua University)
- Xin Zhang (MSRA, Tsinghua University)
- Kerui Min (UIUC, now BosonNLP)
- Zhihan Zhou (UIUC, now PennState)
- Hossein Mobahi (UIUC, now Google)
- Guangcan Liu (UIUC, now NUIST)
- Xiaodong Li (Stanford, now UC Davis)
- Carlos Fernandez (Stanford, now NYU)






THANK YOU!

Questions, please?

 $\circ\, \tau =$  $+$ 

$$D \circ \tau \;=\; A + E \qquad \min \;\; \|A\|_* + \lambda \|E\|_1$$