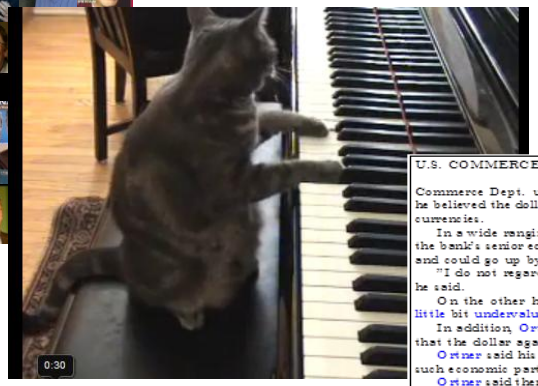
Effective Representations for High-Dimensional Visual Data

John Wright
Electrical Engineering
Columbia University

Datasets are massive, high-dimensional…

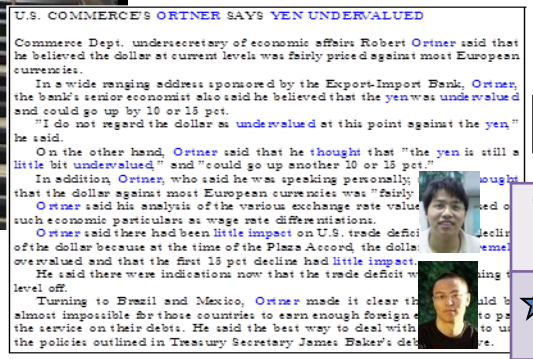
Images

> 1M pixels



Videos

> 1B voxels


U.S. COMMERCE'S ORTNER SAYS YEN UNDERVALUED

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

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

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

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

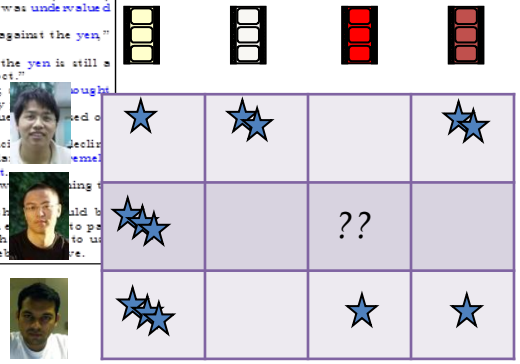
In addition, Ortner, who said he was speaking personally, that the dollar against most European currencies was "fairly

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

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

He said there were indications now that the trade deficit w level off.

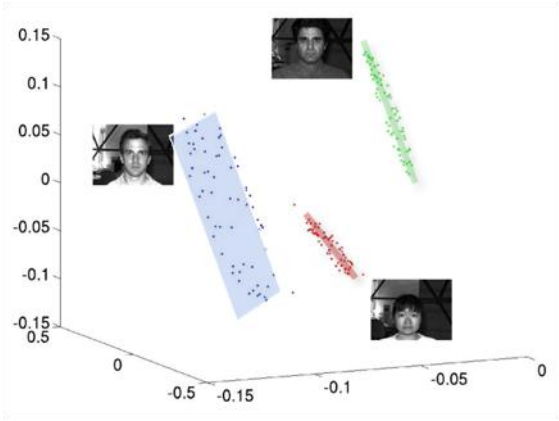
Turning to Brazil and Mexico, Ortner made it clear th almost impossible for those countries to earn enough foreign the service on their debts. He said the best way to deal with the policies outlined in Treasury Secretary James Baker's de

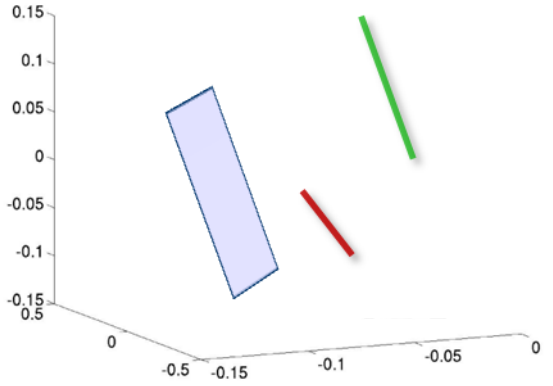


Web data

> 10B+ websites

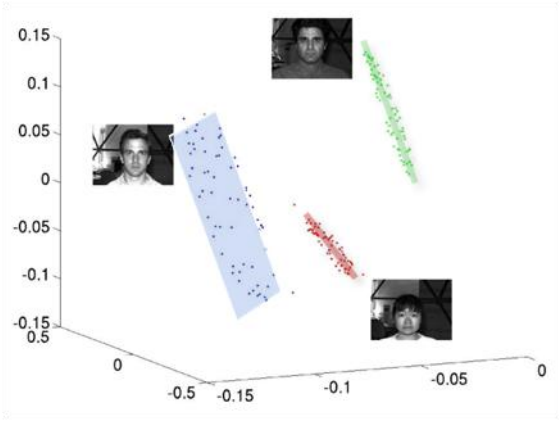
… intrinsic structures are low-dimensional

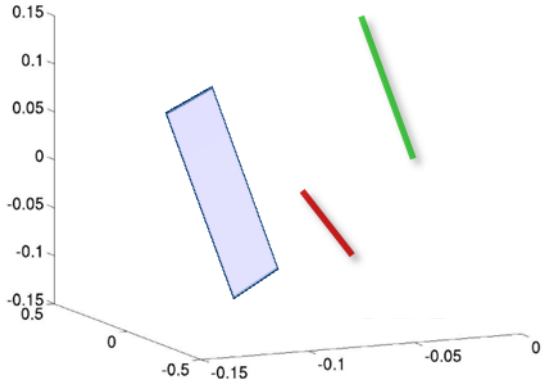




How can we model ***low-dimensional structure*** *in* ***high-dimensional data?***

… intrinsic structures are low-dimensional





*How can we model **low-dimensional structure** in **high-dimensional data?***

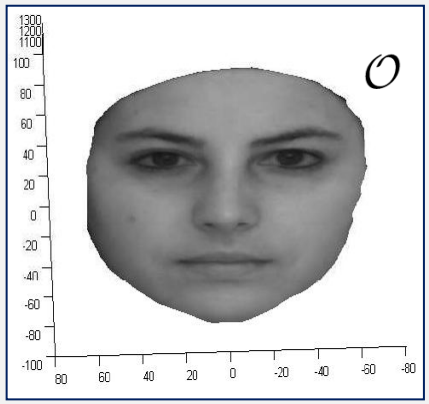
1. *Analytically?*
2. *Automatically?*

… and what does this buy us for engineering applications?

The space/model of images of an object?

Conceptual Problem:

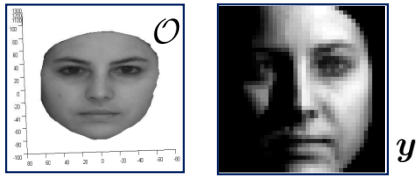
*Given (information about) an object $\mathcal{O}$, can we provably **detect** or **recognize** the object from a new image $\boldsymbol{y}$?*



This talk:

*Attempt this for **fixed pose**, variations in **illumination,** (and possibly occlusion).*

The goal – practical?

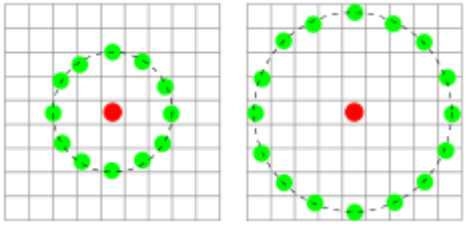


Many **engineering approaches** to mitigate illumination:

Quotient images
[Riklin-Raviv, Shashua '99] [Wang, Li, Wang '04]


Wang, Li, Wang '04

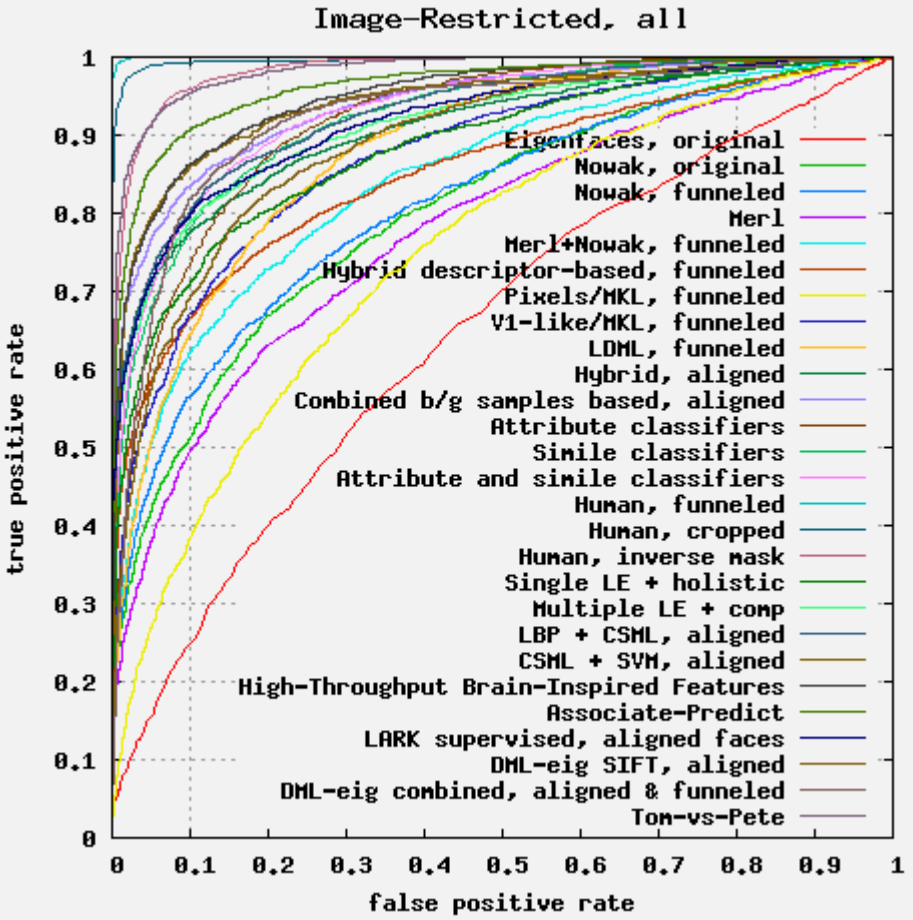
Nonlinear features
SIFT [Lowe '99], HOG [Dalal, Triggs '05]
LBP [Ojala, Pietikäinen, Harwood '94,
Ahonen, Hadid, Pietikainen '06]


Ahonen, Hadid, Pietikainen '06

Total variation minimization
[Yin et. Al.]

Or use **heuristic / intuitive physics** …
[Murase, Nayar '96], [Belhumeur , Kriegman '98], [Basri, Jacobs '03],
[Ramamoorthi '01], [Basri, Frolova '04], [W. et. al. '09], many others…

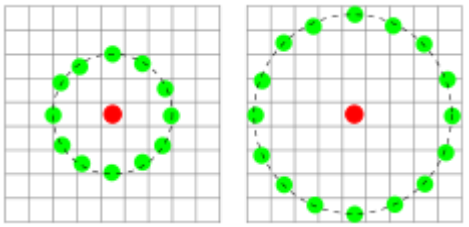
The goal – practical?



Labeled Faces in the Wild results
http://vis-www.cs.umass.edu/lfw/results.html

nation:

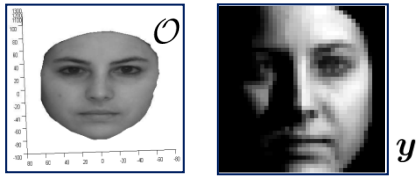
Wang, Li, Wang '04



Ahonen, Hadid, Pietikainen '06

Jacobs '03],
ny others…

The goal – practical?

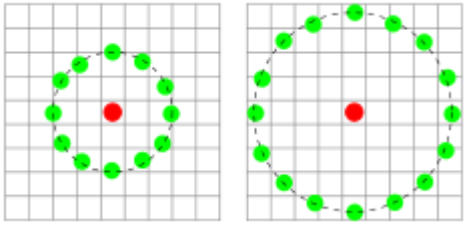


Many **engineering approaches** to mitigate illumination:

Quotient images
[Riklin-Raviv, Shashua '99] [Wang, Li, Wang '04]


Wang, Li, Wang '04

Nonlinear features
SIFT [Lowe '99], HOG [Dalal, Triggs '05]
LBP [Ojala, Pietikäinen, Harwood '94,
Ahonen, Hadid, Pietikainen '06]

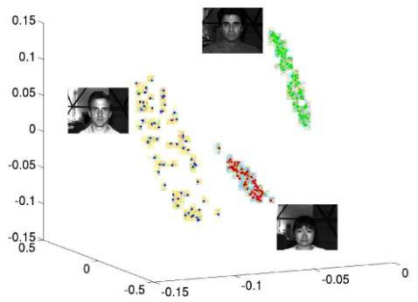

Ahonen, Hadid, Pietikainen '06

Total variation minimization
[Yin et. Al.]

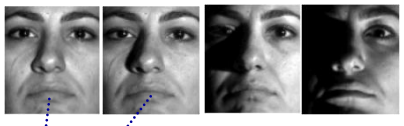
Or use **heuristic / intuitive physics** …
[Murase, Nayar '96], [Belhumeur , Kriegman '98], [Basri, Jacobs '03],
[Ramamoorthi '01], [Basri, Frolova '04], [W. et. al. '09], many others…

Ex: Face recognition with ℓ^1-regression:

Linear subspace model for images of same face under varying illumination:



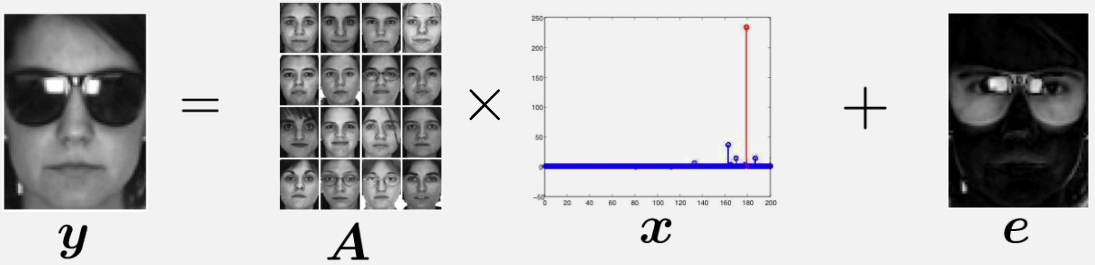
Subject i training



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

If test image $\boldsymbol{y} \in \mathbb{R}^m$ is also of subject i, then $\boldsymbol{y} = A_i \boldsymbol{x_i}$ for some $\boldsymbol{x}_i \in \mathbb{R}^k$.

Given $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{e}_0$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $m \ll n$, recover $\boldsymbol{x}_0$ and $\boldsymbol{e}_0$.



[W., Yang, Ganesh, Sastry, Ma, TPAMI'09]

Ex: Face recognition with ℓ^1-regression:

Underdetermined system of linear equations in unknowns $\boldsymbol{x}, \boldsymbol{e}$:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{e} \end{bmatrix} \doteq \boldsymbol{B}\boldsymbol{w} \qquad \boldsymbol{B} \in \mathbb{R}^{m \times m+n} \quad \boldsymbol{w} \in \mathbb{R}^{m+n}$$

Solution is not unique … but

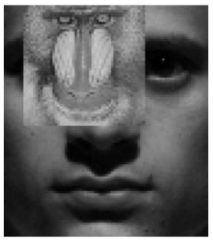
- $\boldsymbol{x}$ should be *sparse*: ideally, only supported on images of the same subject
- $\boldsymbol{e}$ expected to be *sparse*: occlusion only affects a subset of the pixels

Seek the *sparsest* solution:

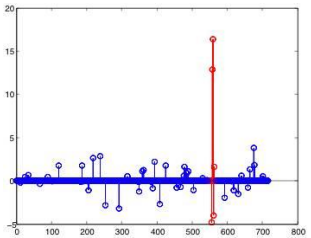
$$\min \|\boldsymbol{x}\|_0 + \|\boldsymbol{e}\|_0 \ \text{ subj } \ \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$$

$$\Downarrow \ \text{convex relaxation} \ \Downarrow$$

$$\min \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \ \text{ subj } \ \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$$



$\boldsymbol{y}$



$\hat{\boldsymbol{x}}_1$

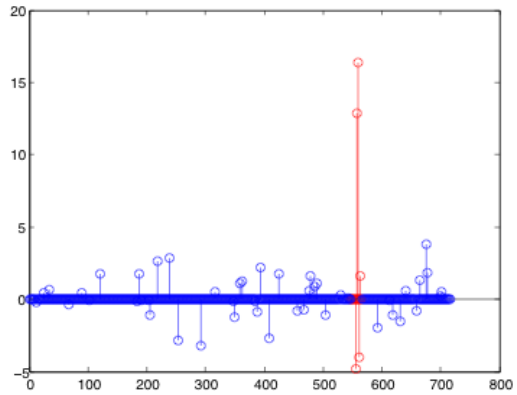
$\hat{\boldsymbol{e}}_1$

$\hat{\boldsymbol{y}} = \boldsymbol{y} - \hat{\boldsymbol{e}}_1$

[W., Yang, Ganesh, Sastry, Ma, TPAMI'09]

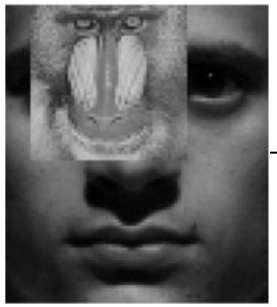
Ex: Face recognition with ℓ^1-regression:

$$\hat{\boldsymbol{x}}_1 \;=\; \arg\min \; \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1.$$



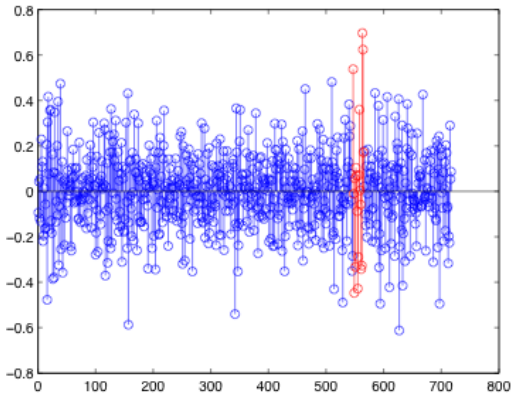
$\hat{\boldsymbol{e}}_1$ $\hat{\boldsymbol{y}}_0 = A\hat{\boldsymbol{x}}_1$

Input: $\boldsymbol{y} \in \mathbb{R}^D$



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

$$\hat{\boldsymbol{x}}_2 \;=\; \arg\min_{\boldsymbol{x}} \; \|\boldsymbol{y} - A\boldsymbol{x}\|_2.$$



$\hat{\boldsymbol{e}}_2$ $\hat{\boldsymbol{y}}_0 = A\hat{\boldsymbol{x}}_2$

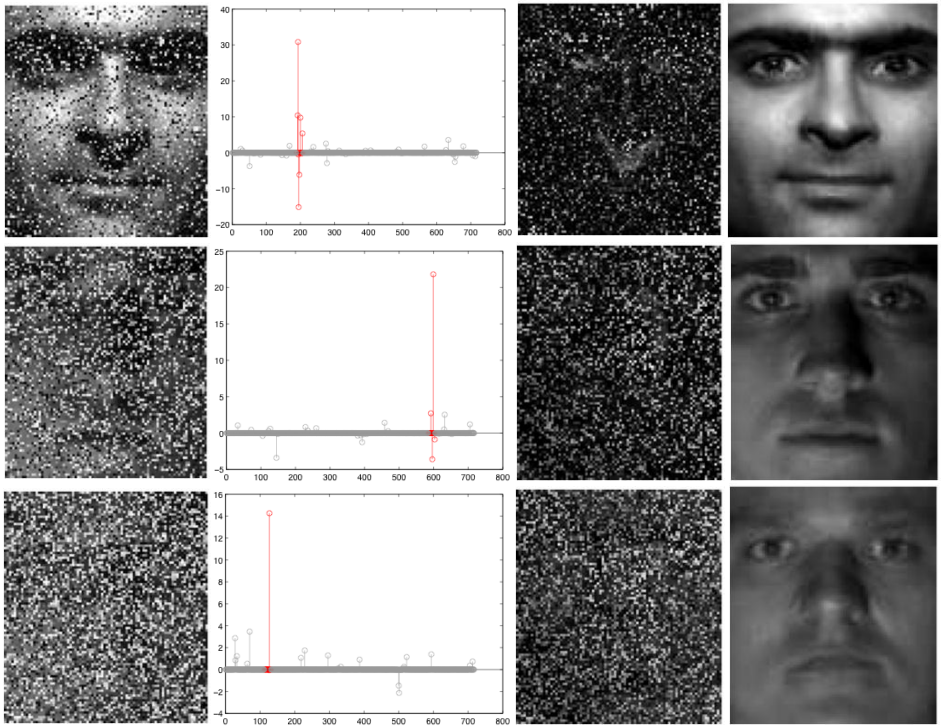
[W., Yang, Ganesh, Sastry, Ma, TPAMI'09]

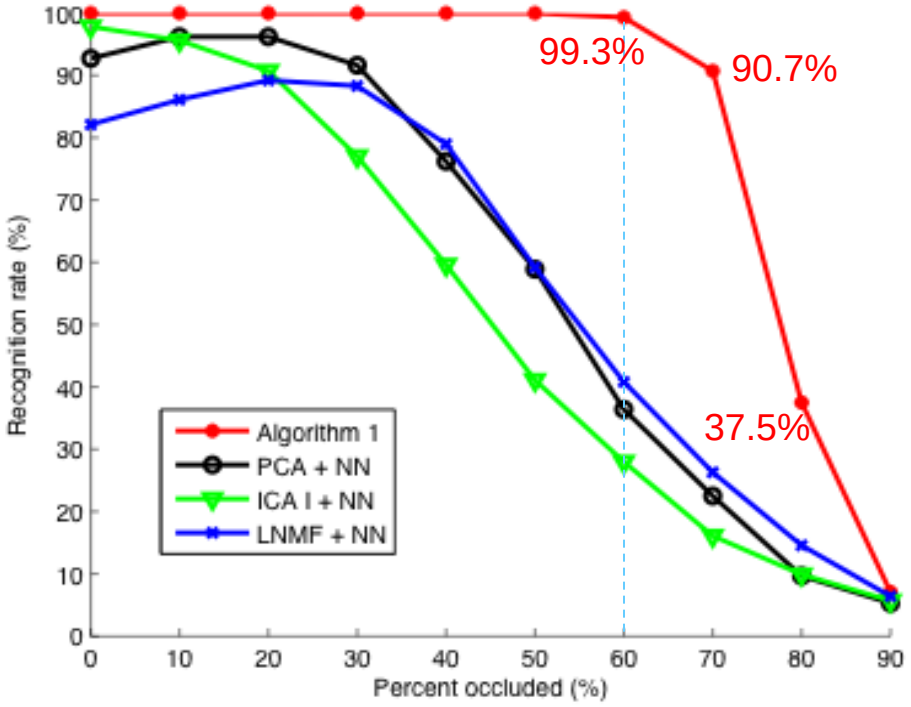
Ex: Face recognition with ℓ^1-regression:

Extended Yale B Database
(38 subjects)

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

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





[W., Yang, Ganesh, Sastry, Ma, TPAMI'09]

Ex: Face recognition with ℓ^1-regression:

Theorem 1 (W. and Ma, '10). *For any $\delta > 0, \rho < 1$, $\exists\, \nu_0(\delta, \rho) > 0$, such that if $\nu < \nu_0$ and $\alpha < \alpha_0(\delta, \nu, \rho)$, then with error support J and signs $\boldsymbol{\sigma}$ chosen uniformly at random,*

$$\lim_{m \to \infty} \mathbb{P}_{A,J,\sigma} \left[\ell^1\text{-minimization recovers all } \alpha m\text{-sparse } \boldsymbol{x} \right] = 1.$$

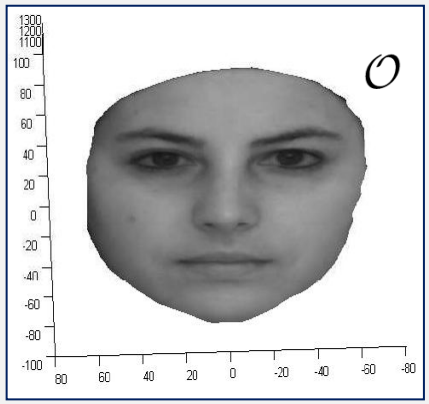
If $\boldsymbol{A}$ is "nice" and the model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ fits, can make strong statements about the performance of ℓ^1 regression.

[W., Ma, Tr. IT '09]
See also … [Candes+Tao '05], [Li '11], [Ngyuen+Tran '12], [McCoy+Tropp '13], [Foygel+Mackey '13]

The space/model of face images of a person?

Conceptual Problem:

*Given (information about) an object $\mathcal{O}$, can we provably **detect** or **recognize** the object from a new image $\boldsymbol{y}$?*



This talk:

*Attempt this for **fixed pose**, variations in **illumination,** (and possibly occlusion).*

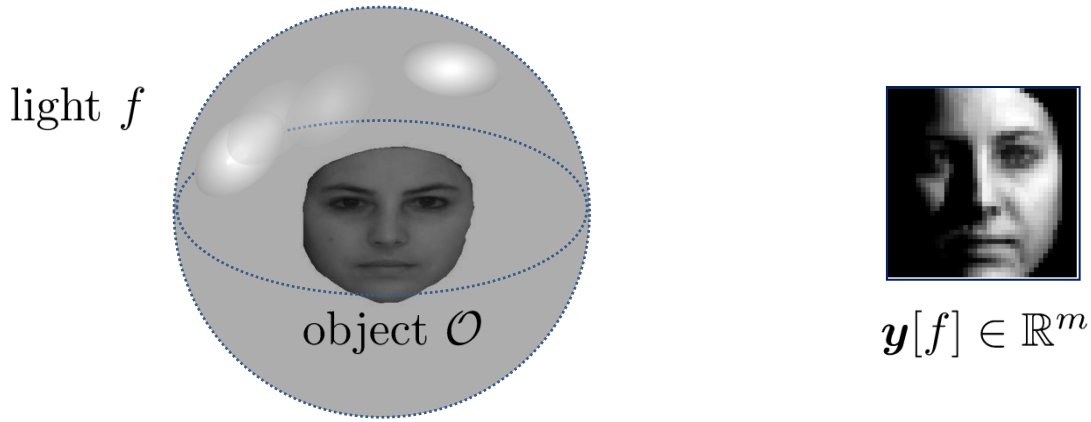
Geometry of illumination variations

Distant illumination identified with a Riemann integrable, nonnegative function $f : \mathbb{S}^2 \to \mathbb{R}_+$.

Assume a **linear sensor response**. The **image** $\boldsymbol{y}[f]$ can often be written as

$$\boldsymbol{y}[f] = \int_{\boldsymbol{u} \in \mathbb{S}^2} f(\boldsymbol{u}) \bar{\boldsymbol{y}}[\boldsymbol{u}] \, d\boldsymbol{u}$$

What is the set of possible images $\boldsymbol{y}[f]$ of $\mathcal{O}$?



Geometry of illumination variations

Distant illumination identified with a Riemann integrable, nonnegative function $f : \mathbb{S}^2 \to \mathbb{R}_+$.

Assume a **linear sensor response**. The **image** $\boldsymbol{y}[f]$ can often be written as

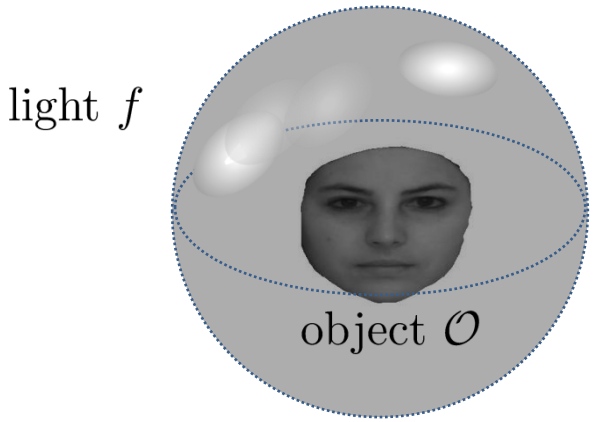
$$\boldsymbol{y}[f] = \int_{\boldsymbol{u} \in \mathbb{S}^2} f(\boldsymbol{u}) \bar{\boldsymbol{y}}[\boldsymbol{u}] \, d\boldsymbol{u}$$

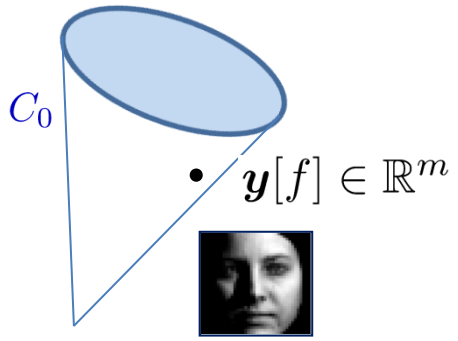
What is the set of possible images $\boldsymbol{y}[f]$ of $\mathcal{O}$?

Let $\mathcal{F}_0 \doteq \{ f : \mathbb{S}^2 \to \mathbb{R}_+ \mid f \text{ Riemann integrable} \}$, *and*

$C_0 = \{ \boldsymbol{y}[f] \mid f \in \mathcal{F}_0 \} \subset \mathbb{R}^m$.

The set C_0 *is a* ***convex cone.*** [Belhumeur + Kriegman '98]



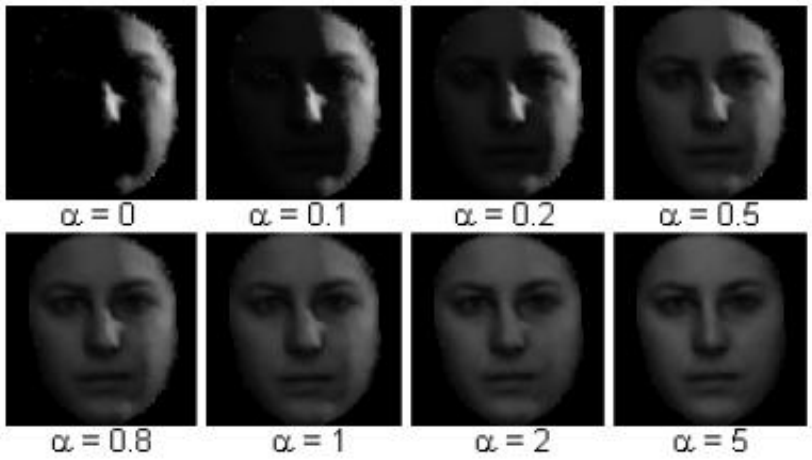


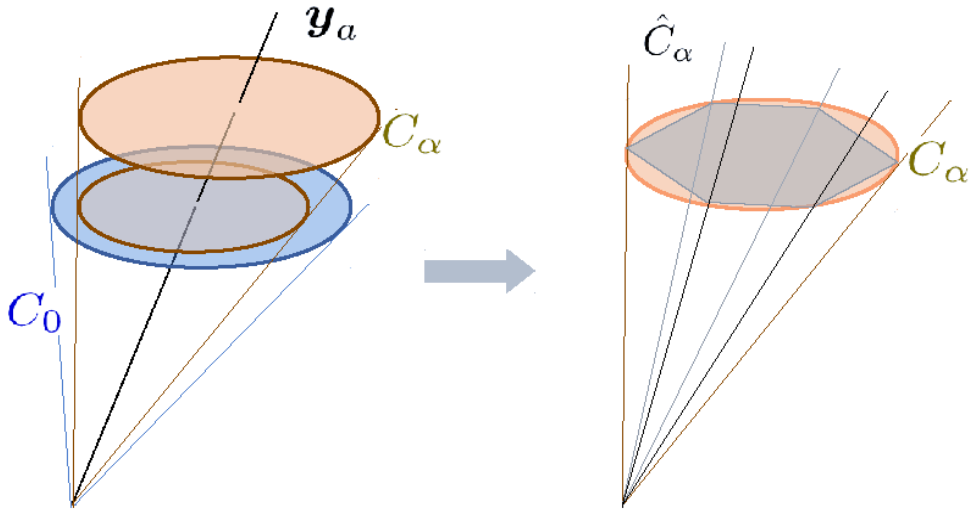
Geometry of illumination variations

Ambient cone model. *Illumination is the sum of ambient and directional (arbitrary) components:*

$$\mathcal{F}_\alpha \doteq \{ f_d + \alpha\omega \mid f_d : \mathbb{S}^2 \to \mathbb{R}_+, \text{Riemann integrable}, \ \|f_d\|_{L^1} \leq 1 \}$$

and corresponding cone of possible images: $C_\alpha \doteq \text{cone}\Big(\boldsymbol{y}[\mathcal{F}_\alpha] \Big)$





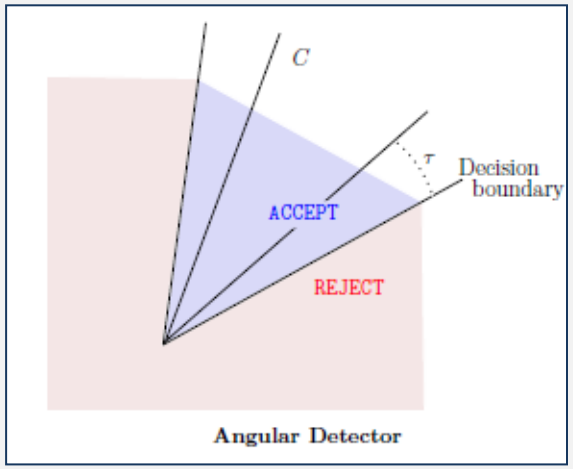
Approximating a convex cone

Angular detector: is $\boldsymbol{y}$ in or close to C_α ?

Accept *any input* $\boldsymbol{y}$ *that is an image of* $\mathcal{O}$ *under some valid illumination.*

Reject *any input* $\boldsymbol{y}$ *that is not.*

$$\mathfrak{D}_\tau[\boldsymbol{y}] = \begin{cases} \text{ACCEPT} & \angle(\boldsymbol{y}, C) \leq \tau \\ \text{REJECT} & \angle(\boldsymbol{y}, C) > \tau \end{cases}$$

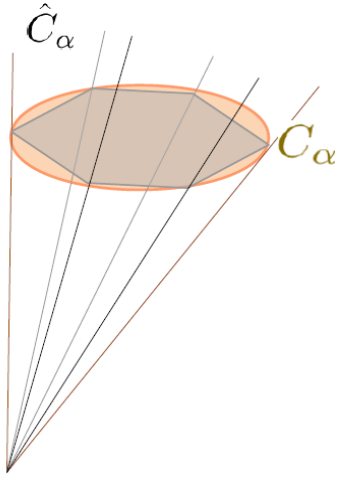


Work with a "Hausdorff distance" between cones:

$$\begin{aligned} \delta(C, C') &\doteq d_{\text{Hausdorff}}\Big(C \cap B(0,1), C' \cap B(0,1)\Big) \\ &= \max\left\{ \sup_{\boldsymbol{y} \in C \cap B(0,1)} d(\boldsymbol{y}, C'), \sup_{\boldsymbol{y}' \in C' \cap B(0,1)} d(\boldsymbol{y}', C) \right\} \end{aligned}$$

$\mathcal{O}$

$\boldsymbol{y}$



If C' ***approximates*** C ***in Hausdorff sense, we lose little in working with*** C'.

Cone Approximation: Existing Results

Approximation of **general** convex bodies in high dimensions is a disaster:

Theorem 2. [Bronstein, Ivanov '76] *Let $K \subset \mathbb{R}^m$ be a convex body. There exists an ε-approximation to K in Hausdorff distance, with $O\left((\mathrm{diam}(K)/\varepsilon)^{\frac{m-1}{2}}\right)$ vertices. For the unit sphere, this is optimal up to a constant.*

In vision literature, results for **average case**, for **convex objects**:

Informal claim [Basri and Jacobs '03, Basri and Frolova '04]:
For an convex, Lambertian object $\mathcal{O}$, there exists a subspace Γ, with $\dim(\Gamma) = 9$, such that if $\boldsymbol{u} \sim \mathrm{uni}\left(\mathbb{S}^2\right)$ is a uniformly oriented random point source,

$$\frac{\mathbb{E}\left[\|\bar{\boldsymbol{y}}[\boldsymbol{u}] - \mathcal{P}_\Gamma \bar{\boldsymbol{y}}[\boldsymbol{u}]\|_2^2\right]}{\mathbb{E}\left[\|\bar{\boldsymbol{y}}[\boldsymbol{u}]\|_2^2\right]} \leq .02$$

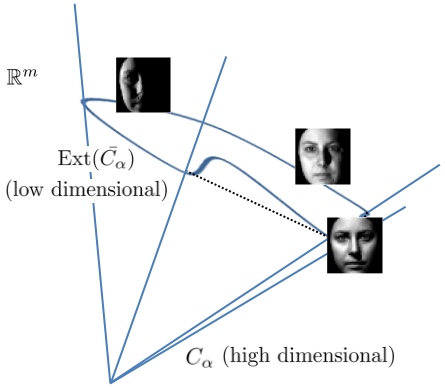
Is there any special structure we can use here?

Extreme rays

Image formation: $\boldsymbol{y}[f] = \int_{\boldsymbol{u} \in \mathbb{S}^2} f(\boldsymbol{u}) \bar{\boldsymbol{y}}[\boldsymbol{u}] \, d\boldsymbol{u}$.

$\bar{\boldsymbol{y}}[\boldsymbol{u}] \in \mathbb{R}^m$ are images under **directional illumination**:

Extreme rays lie on a low dimensional submanifold of $\mathbb{R}^m$ ***:***

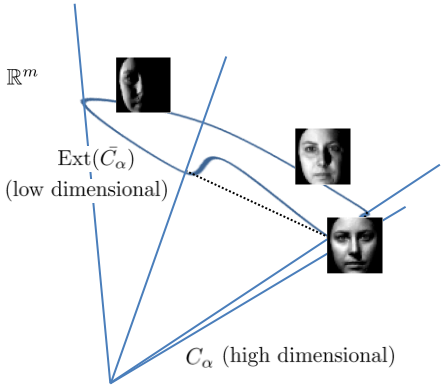


Extreme rays

Image formation: $\boldsymbol{y}[f] = \int_{\boldsymbol{u} \in \mathbb{S}^2} f(\boldsymbol{u}) \bar{\boldsymbol{y}}[\boldsymbol{u}] \, d\boldsymbol{u}$.

$\bar{\boldsymbol{y}}[\boldsymbol{u}] \in \mathbb{R}^m$ are images under **directional illumination**:

Extreme rays lie on a low dimensional submanifold of $\mathbb{R}^m$ ***:***



Lemma. *[Just approximate the extreme rays]. Let* $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N \in \mathbb{S}^2$ *. Then*

$$\delta\Big(C_\alpha, \ \mathrm{cone}\,\{\bar{\boldsymbol{y}}[\boldsymbol{u}_i] + \alpha \boldsymbol{y}_a\} \Big) \ \leq \ \frac{2}{\eta_\star \alpha \|\boldsymbol{y}_a\|} \times \sup_{\boldsymbol{u}} \min_i \|\bar{\boldsymbol{y}}[\boldsymbol{u}] - \bar{\boldsymbol{y}}[\boldsymbol{u}_i]\|_2,$$

where $\eta_\star = \sup_{\|\boldsymbol{w}\|_2 \leq 1} \min_i \left\langle \boldsymbol{w}, \frac{\bar{\boldsymbol{y}}[\boldsymbol{u}_i]}{\|\bar{\boldsymbol{y}}[\boldsymbol{u}_i]\|_2} \right\rangle \ \geq \ 1/\sqrt{m}.$

Cone Approximation: New Results

Theorem 3 (Zhang, Mu, Kuo, W. '12) *(sketch) Under our previously stated hypotheses, for all* $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{S}^2$,

$$\|\bar{\boldsymbol{y}}[\boldsymbol{u}] - \bar{\boldsymbol{y}}[\boldsymbol{u}']\|_2 \;\leq\; \frac{C_{sensor}}{1-\nu_\star} \times \left(\operatorname{area}(\partial\mathcal{O})\, \|\boldsymbol{u}-\boldsymbol{u}'\|_2^2 + \chi_\star \operatorname{diam}(\mathcal{O})\, \|\boldsymbol{u}-\boldsymbol{u}'\|_2 \right)^{1/2}$$

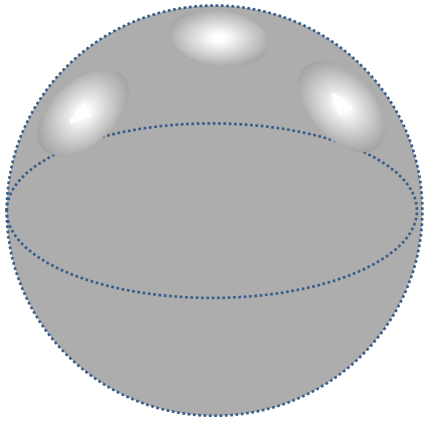
where C_{sensor} *depends on the parameters of the imaging system.*

If we define an **illumination covering number**

$$N(C,\gamma) \;=\; \min\{\#V \mid \delta(C, \operatorname{cone}(V)) \;\leq\; \gamma\},$$

This implies in general, $N(C_\alpha,\gamma) \;\leq\; \dfrac{h(\mathcal{O},\text{sensor})}{(\alpha\gamma)^4}$,

and for convex objects $N(C_\alpha,\gamma) \;\leq\; \dfrac{h(\mathcal{O},\text{sensor})}{(\alpha\gamma)^2}$.

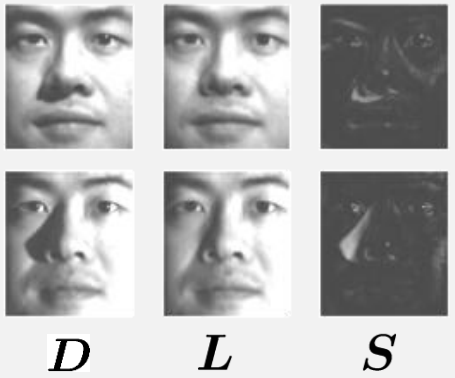


Proof Sketch

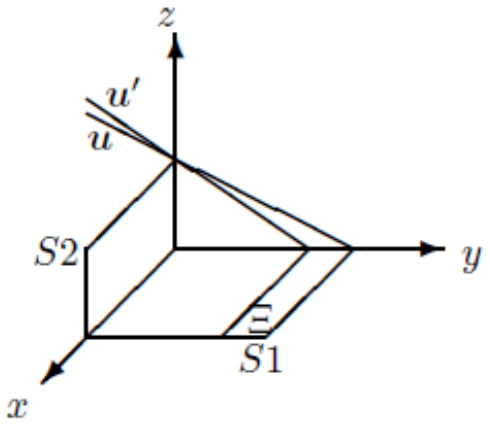
Control the complexity of cone approximation in terms of object convexity defect.

The conceptual idea:

Separate the direct illumination operator into a **low-rank** part (due to smooth variations) and a **sparse** part (due to cast shadows):



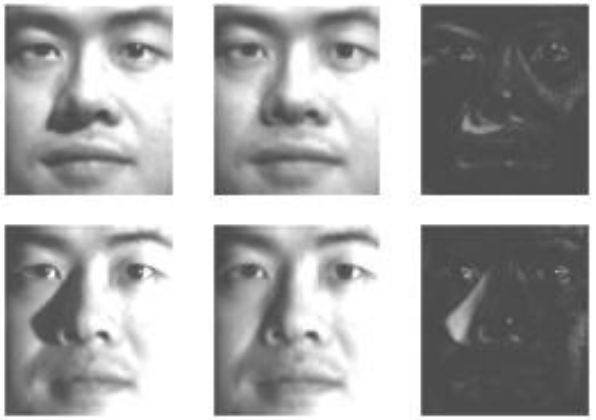
The low-rank term can be bounded by direct calculation; the sparse term needs a more involved accounting.



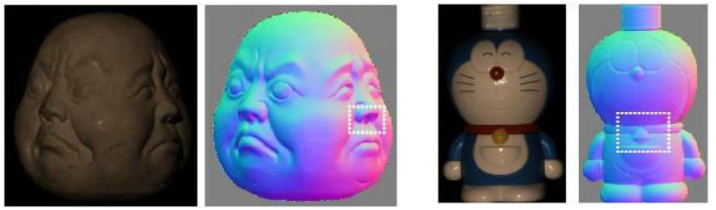
LR+S captures physical intuition

Calculations with the Lambertian sphere suggest that sets of images of **smooth, near-convex** objects should be approximately **low-rank**:
[Basri+Jacobs '03,Ramamoorthi '04].

Cast shadows are often **sparse**
[W., Yang, … Ma, '09] , [Candes, Li, Ma, W. '11]:



Observations used in photometric stereo:
[Wu et. Al. '11] :

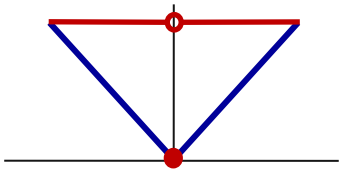


How can we exploit these structures to obtain compact representation/approximation of the cone?

Low-rank and sparse recovery

Convex relaxation: $\min \ \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \quad \text{s.t} \quad \boldsymbol{L} + \boldsymbol{S} = \boldsymbol{D}.$

[Chadrasekaran et. al. '11]



Provably effective for recovery (and **statistical estimation**):

Theorem 5 (Candès, Li, Ma, W. '11). *If* $\boldsymbol{L}_0 \in \mathbb{R}^{m \times n}$, $m \geq n$ *has rank*

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

and $\boldsymbol{S}_0$ *has Bernoulli support with error probability* $\rho \leq \rho_s^\star$, *then w.h.p.,*

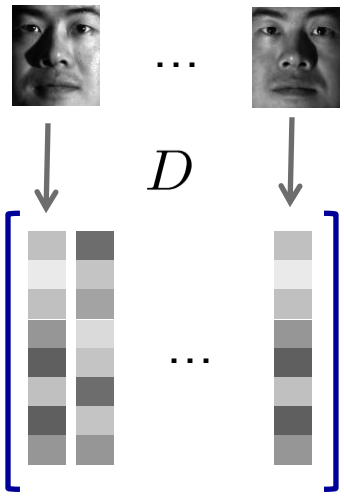
$$(\boldsymbol{L}_0, \boldsymbol{S}_0) \ = \ \arg\min \|\boldsymbol{L}\|_* + \frac{1}{\sqrt{m}} \|\boldsymbol{S}\|_1 \quad \text{subj} \quad \boldsymbol{L} + \boldsymbol{S} = \boldsymbol{L}_0 + \boldsymbol{S}_0,$$

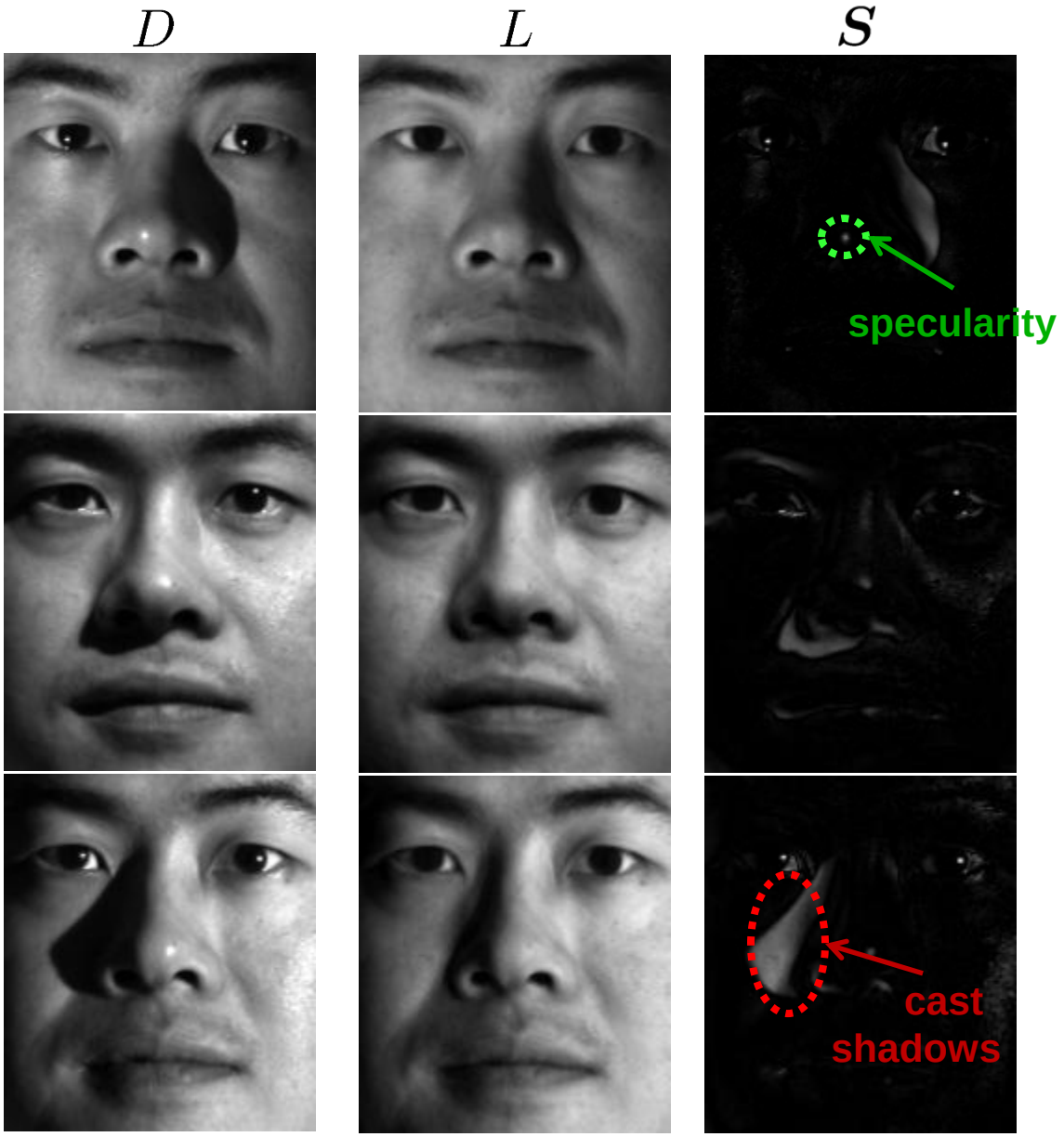
and the minimizer is unique.

See also … [Chadrasekaran,Sanghavi,Willsky,Parillo'11], [Candes, Li, Ma, W. '11], [Hsu, Kakade, Zhang '12], [Agarwal, Wainwright '12],

Recover low-rank and sparse components

58 images of one person
under varying lighting:





Cone-preserving dimensionality reduction

A low-rank and sparse decomposition that provably **preserves verification performance**?

Would need to solve

$$\text{minimize } \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1 \quad \text{s.t.} \quad \delta\Big(\text{cone}(\boldsymbol{L}+\boldsymbol{S}), \text{cone}(\boldsymbol{A})\Big) \;\leq\; \gamma$$

… but constraint is very complicated. Relax to:

Theorem 6 (Zhang, Mu, Kuo, W. '13) *Let* $(\boldsymbol{L}_\star, \boldsymbol{S}_\star)$ *solve*

$$\text{minimize } \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1 \quad \text{s.t.} \quad \begin{bmatrix} \boldsymbol{I} & \boldsymbol{L}+\boldsymbol{S}-\boldsymbol{A} \\ \boldsymbol{L}^*+\boldsymbol{S}^*-\boldsymbol{A}^* & \gamma'\boldsymbol{A}^*\boldsymbol{A}-\boldsymbol{\mu} \end{bmatrix} \succeq \boldsymbol{0},\; \boldsymbol{\mu} \geq \boldsymbol{0},$$

with $\gamma' = \frac{\gamma}{1+\gamma}$. *Then* $\delta\Big(\text{cone}(\boldsymbol{L}_\star+\boldsymbol{S}_\star), \text{cone}(\boldsymbol{A})\Big) \;\leq\; \gamma.$

Provable Low-Dim Structures of Visual Data

Guaranteed Illumination Models for Nonconvex Objects, Zhang, Mu, Kuo, W., Arxiv '13

Compressive Principal Component Pursuit, W., Ganesh, Min, Ma, I&I '13

Towards a Practical Automatic Face Recognition, Wagner, W., Ganesh, Zhou, Mobahi, Ma, PAMI '12

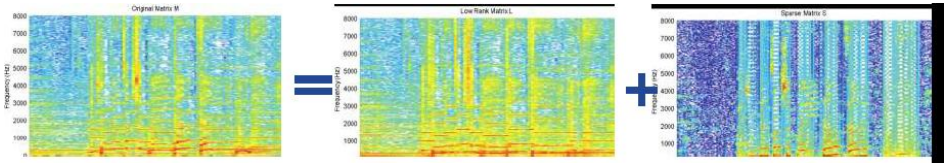
Robust Principal Component Analysis? Candes, Li, Ma, W. JACM '11

Dense Error Correction via ℓ^1 Minimization. W. and Ma, Information Theory '10

Robust Face Recognition via Sparse Representation, W., Yang, Ganesh,, Sastry, Ma, PAMI '09

Low-rank and sparse: applications

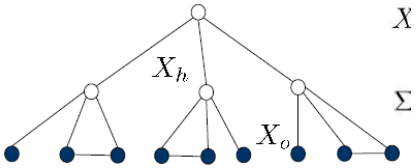
Audio: music and voices
[Huang et. al. ’12]



Video: foreground/background
[CLMW ’11]

Alignment: images online
[PGWM’11]

Learning graphical models:
[Chandrasekaran et. al. ’12]



Many others … text (robust topic models), community detection in graphs, subspace segmentation, 3D reconstruction…

Generalizations

Upgrade a certificate to a *compressive* certificate. More generally:

Multiple structure decomposition:

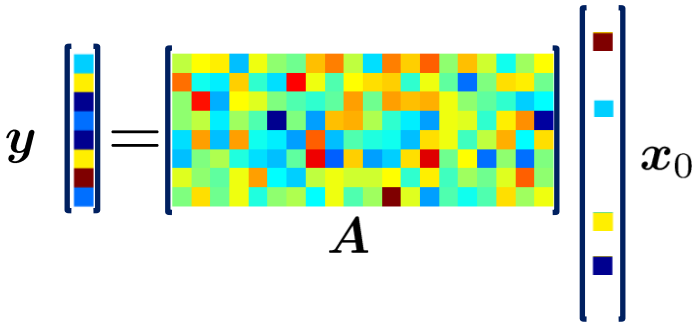
$$\min \textstyle\sum_i \lambda_i \|\boldsymbol{X}_i\|_{\diamond_i} \quad \text{s.t.} \quad \sum_i \boldsymbol{X}_i = \boldsymbol{D}$$

Compressive multiple structure decomposition:

$$\min \textstyle\sum_i \lambda_i \|\boldsymbol{X}_i\|_{\diamond_i} \quad \text{s.t.} \quad \mathcal{P}_Q[\sum_i \boldsymbol{X}_i] = \mathcal{P}_Q[\boldsymbol{D}]$$

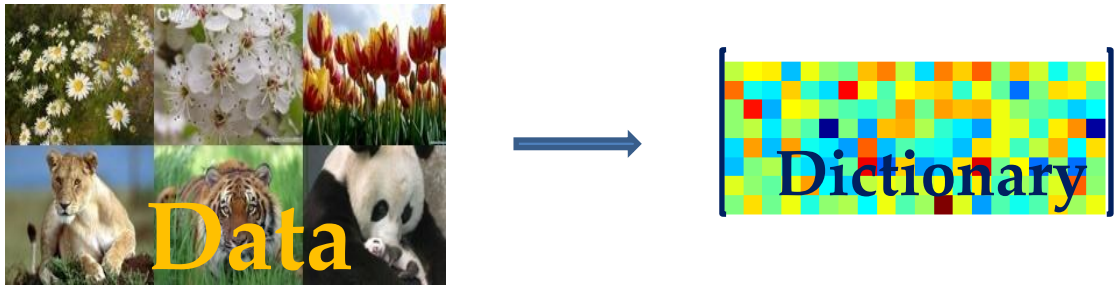
Examples: **PCP** [Chadrasekaran,Sanghavi,Willsky,Parillo'11,CLMW '11],
outlier pursuit [Xu+Caramanis+Sanghavi],
morphological component analysis [Bobin et. al.], many more …

Learning simple signal models?



with $x_0 \in \mathbb{R}^n$ **sparse** - most of the $x_0(i)$ are zero.

Good model for many types of imagery data, especially if we can **learn the dictionary** $A = [a_1 \mid \cdots \mid a_n] \in \mathbb{R}^{m \times n}$

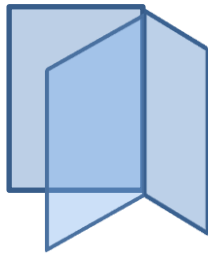


The model problem

Given $\boldsymbol{Y} \approx \boldsymbol{A}\boldsymbol{X}$ with $\boldsymbol{x}_j$ sparse,
$(\boldsymbol{A}, \boldsymbol{X})$ unknown, recover $\boldsymbol{A}$ and $\boldsymbol{X}$.

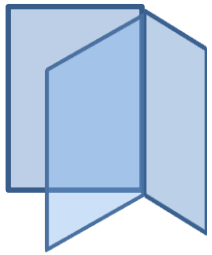
Ambiguities: $(\boldsymbol{A}, \boldsymbol{X})$ or $(\boldsymbol{A}\boldsymbol{\Pi}\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{-1}\boldsymbol{\Pi}^{*}\boldsymbol{X})$?

Peculiar geometry:

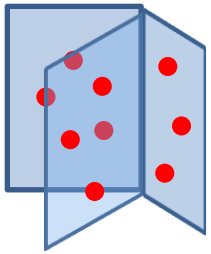


k column subspaces of $\boldsymbol{A}$

When is dictionary learning well-posed?



$\binom{n}{k}$ k column subspaces of $\boldsymbol{A}$

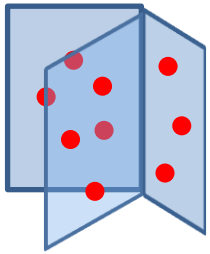


$\boldsymbol{Y}$ k+1 points per subspace

Solution is unique:

Theorem 1 (ess. Aharon et. al. '05) *(sketch) There exists k column sparse $\boldsymbol{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_p]$, of size $p = (k+1)\binom{n}{k}$ such that if we observe $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, $(\boldsymbol{A}, \boldsymbol{X})$ is essentially the only k-column sparse factorization of $\boldsymbol{Y}$.*

When does a learned dictionary generalize?



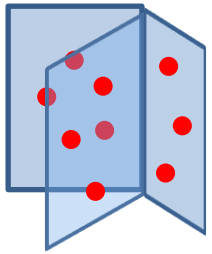
Theorem 2 (Vainsencher, Mannor and Bruckstein '11) *(sketch) If* $\boldsymbol{y} \sim_{iid} \mu$ *on* $\mathbb{S}^{m-1}$, $p > p_0$, $\lambda > \lambda_0$, *then with prob.* $1 - e^{-t}$ *in* $\boldsymbol{Y}$,

$$\mathbb{E}_{\boldsymbol{y}} \min_{\|\boldsymbol{x}\|_1 \leq \lambda} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|$$

$$\leq \frac{1.1}{p} \sum_i \|\boldsymbol{y}_i - \boldsymbol{A}\boldsymbol{x}_i\| + 9\frac{mn\log(\lambda p)+t}{p}$$

See also [Maurer and Pontil '10].

How can we learn a good dictionary?



$$\boldsymbol{Y} \approx \boldsymbol{A}\boldsymbol{X},\ \boldsymbol{X} \text{ sparse.}$$

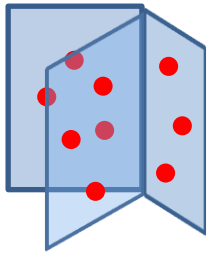
Alternating directions to minimize sparsity surrogate

[Engan et. al., '99, Aharon et. al. '05, Yaghoobi '10]

$$\min \tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}\|_F^2 + J(\boldsymbol{X})$$

Recently: Supervised variants [Mairal et. al. '08], structured dictionaries [Rubenstein et. al. '10], highly scalable variants [Mairal et. al. '10] … and many, many more…

Is the desired solution a local minimum?



$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X},\ \boldsymbol{X} \text{ sparse.}$$

$$\min \quad \|\boldsymbol{X}'\|_1 \quad \text{s.t.} \quad \boldsymbol{Y} = \boldsymbol{A}'\boldsymbol{X}',\ \boldsymbol{A}' \in \mathcal{A}$$

For square $\boldsymbol{A}$, under probabilistic assumptions on $\boldsymbol{X}$, $(\boldsymbol{A}, \boldsymbol{X})$ is a local minimum whp:

Theorem 3 (Gribonval + Schnass '10) *(sketch) Let* $\boldsymbol{X}_{ij} = \boldsymbol{\Omega}_{ij}\boldsymbol{V}_{ij}$, *with* $\boldsymbol{\Omega} \sim \mathrm{Ber}(\theta)$, $\boldsymbol{V} \sim \mathcal{N}(0,1)$. *For square, incoherent* $\boldsymbol{A}$, $(\boldsymbol{A}, \boldsymbol{X})$ *is a local minimum of* $\|\cdot\|_1$ *with high probability, provided* $p = \Omega(n \log n/\theta)$.

Is the desired solution a local minimum?

$$\min \quad \|\boldsymbol{X}'\|_1 \quad \text{s.t.} \quad \boldsymbol{Y} = \boldsymbol{A}'\boldsymbol{X}', \;\; \boldsymbol{A}' \in \mathcal{A}$$

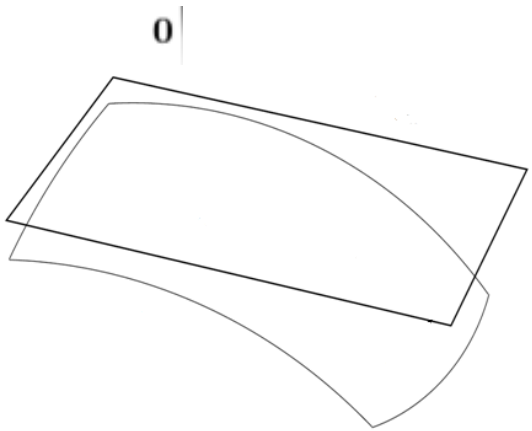
For general $\boldsymbol{A}$, under probabilistic assumptions on $\boldsymbol{X}$, $(\boldsymbol{A}, \boldsymbol{X})$ is a **local minimum whp**:

Theorem 4 (Geng, W., '11). *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$, $k < C/\mu(\mathbf{A})$, *and* $\mathbf{X} \in \mathbb{R}^{n\times p}$ *with random* k*-sparse support, independent Gaussian nonzeros. Then* $(\mathbf{A}, \mathbf{X})$ *is a local minimum of the* ℓ^1*-norm wp* $\geq 1 - \tilde{O}(n^{3/2}k^{1/2}p^{-1/2})$.

Is this obvious?

Maybe … but surprisingly resistant to analysis …

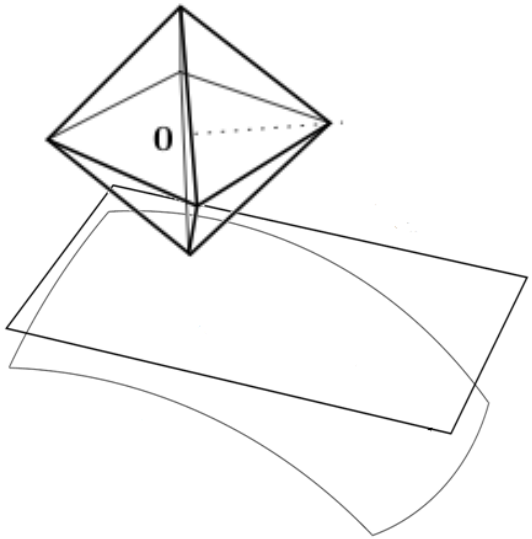
$$\min \quad \|\boldsymbol{X}'\|_1 \quad \text{s.t.} \quad \boldsymbol{Y} = \boldsymbol{A}'\boldsymbol{X}', \;\; \boldsymbol{A}' \in \mathcal{A}$$



Is this obvious?

Maybe … but surprisingly resistant to analysis …

$$\min \quad \|\boldsymbol{X}'\|_1 \quad \text{s.t.} \quad \boldsymbol{Y} = \boldsymbol{A}'\boldsymbol{X}', \;\; \boldsymbol{A}' \in \mathcal{A}$$

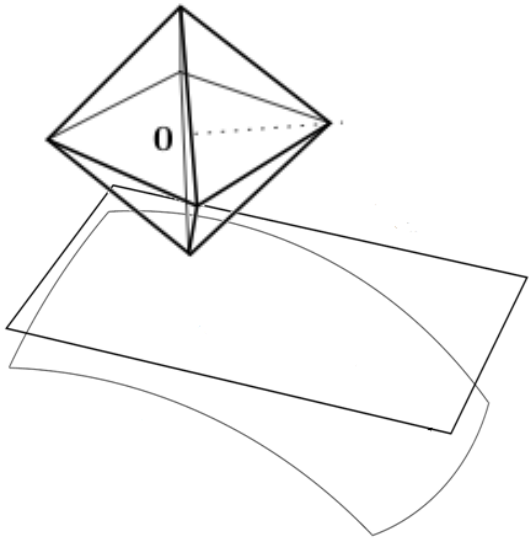


Feasible $(\boldsymbol{A}, \boldsymbol{X})$

Is this obvious?

Maybe … but surprisingly resistant to analysis …

$$\min \quad \|\boldsymbol{X}'\|_1 \quad \text{s.t.} \quad \boldsymbol{Y} = \boldsymbol{A}'\boldsymbol{X}', \;\; \boldsymbol{A}' \in \mathcal{A}$$



Feasible $(\boldsymbol{A}, \boldsymbol{X})$

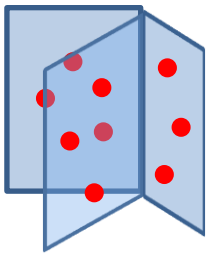
Have to analyze an ℓ^1 problem over an affine space.

RIP ect., fail here
ess. sign-permutation ambiguity

Use ideas from **low-rank recovery**
[Gross '09], [Candes, Li, Ma, W. '12].

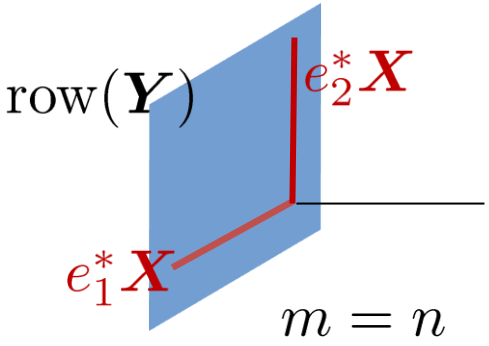
Can we obtain global results?

In general, not easy… special geometry for square case



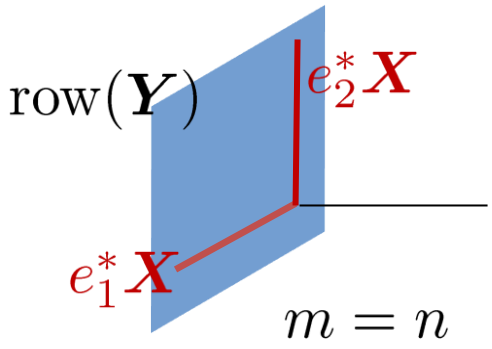
$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X},\ \boldsymbol{X} \text{ sparse.}$$

If $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ nonsingular, $\operatorname{row}(\boldsymbol{X}) = \operatorname{row}(\boldsymbol{Y})$.



Rows of $\boldsymbol{X}$ are sparse vectors in a known subspace.

Uniqueness – square dictionaries

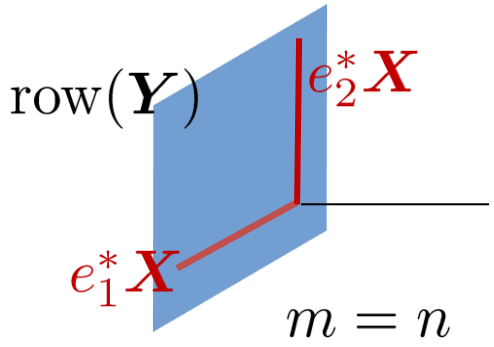


Rows of X are sparse vectors in a known subspace.

If $p > cn \log n$ then whp. rows of X are the sparsest vectors in $\mathrm{row}(Y)$

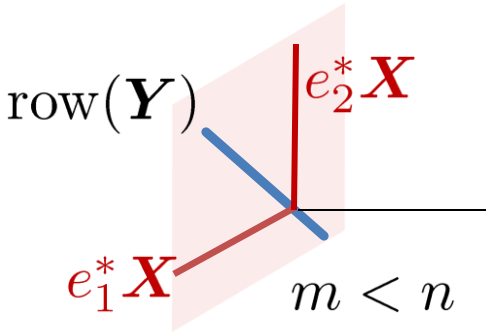
Uniqueness – square dictionaries

Square:



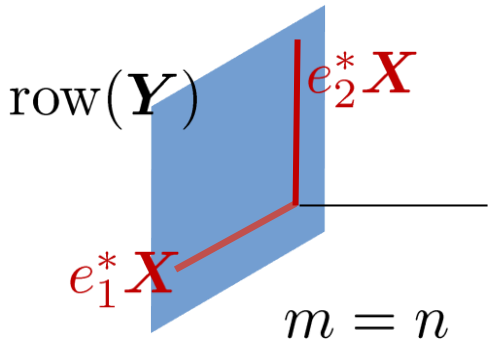
Theorem [Spielman, Wang, W. '11]:
Decomposition essentially unique from $\Omega(n \log n)$ random observations.

Overcomplete:



Theorem [Aharon, Elad, Bruckstein '05]:
Decomposition is essentially unique from $(k+1)\binom{n}{k}$ strategically located observations.

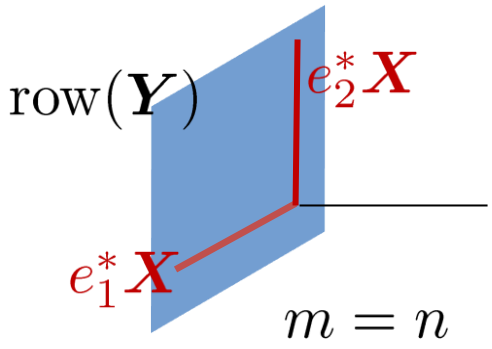
Algorithms – square dictionaries



Rows of $\boldsymbol{X}$ are sparsest vectors in $\mathrm{row}(\boldsymbol{Y})$

$$\text{minimize } \|\boldsymbol{w}^* \boldsymbol{Y}\|_0 \quad \text{subject to} \quad \boldsymbol{w} \neq 0.$$

Algorithms – square dictionaries



Rows of $\boldsymbol{X}$ are sparsest vectors in $\mathrm{row}(\boldsymbol{Y})$

$$\text{minimize } \|\boldsymbol{w}^*\boldsymbol{Y}\|_0 \quad \text{subject to} \quad \boldsymbol{w} \neq 0.$$

$$\text{minimize } \|\boldsymbol{w}^*\boldsymbol{Y}\|_1 \quad \text{subject to} \quad \boldsymbol{r}^*\boldsymbol{w} = 1.$$

Algorithms – square dictionaries

$$\text{minimize } \|\boldsymbol{w}^*\boldsymbol{Y}\|_1 \quad \text{subject to} \quad \boldsymbol{r}^*\boldsymbol{w} = 1.$$

What choice of $\boldsymbol{r}$ will make $\hat{\boldsymbol{w}}^*\boldsymbol{Y} = \boldsymbol{e}_i^*\boldsymbol{X}$?

Change variables $\boldsymbol{q} = \boldsymbol{A}^*\boldsymbol{w}$:

$$\text{minimize } \|\boldsymbol{q}^*\boldsymbol{X}\|_1 \quad \text{subject to} \quad (\boldsymbol{A}^{-1}\boldsymbol{r})^*\boldsymbol{q} = 1.$$

If $\boldsymbol{r} = \boldsymbol{A}\boldsymbol{e}_i$ we're golden …

Don't have this; use $\boldsymbol{y}_j = \sum_{i \in I} X_{ij} \boldsymbol{A}\boldsymbol{e}_i$

Algorithms – square dictionaries

ER-SpUD(SC): Exact Recovery of Sparsely-Used Dictionaries using single columns of Y as constraint vectors.

For $j = 1 \dots p$

Solve $\min_w \ \|w^T Y\|_1$ subject to $(Y e_j)^T w = 1$, and set $s_j = w^T Y$.

Greedy: A Greedy Algorithm to Reconstruct X and A.

1. **REQUIRE:** $\mathcal{S} = \{s_1, \dots, s_T\} \subset \mathbb{R}^p$.
2. For $i = 1 \dots n$

   REPEAT

   $l \leftarrow \arg\min_{s_l \in \mathcal{S}} \|s_l\|_0$, breaking ties arbitrarily

   $x_i = s_l$

   $\mathcal{S} = \mathcal{S} \backslash \{s_l\}$

   UNTIL rank$([x_1, \dots, x_i]) = i$
3. Set $X = [x_1, \dots, x_n]^T$, and $A = YY^T(XY^T)^{-1}$.

Algorithms – square dictionaries

ER-SpUD(DC): Exact Recovery of Sparsely-Used Dictionaries using the sum of two columns of Y as constraint vectors.

1. Randomly pair columns of $\boldsymbol{Y}$ into $p/2$ groups $g_i = \{\boldsymbol{Y}\boldsymbol{e}_{i1}, \boldsymbol{Y}\boldsymbol{e}_{i2}\}$.
2. For $j = 1 \dots p/2$

   Let $\boldsymbol{r}_j = \boldsymbol{Y}\boldsymbol{e}_{j1} + \boldsymbol{Y}\boldsymbol{e}_{j2}$, where $\boldsymbol{Y}\boldsymbol{e}_{j1}, \boldsymbol{Y}\boldsymbol{e}_{j2} \in g_j$.

   Solve $\min_w \|\boldsymbol{w}^T\boldsymbol{Y}\|_1$ subject to $\boldsymbol{r}_j^T\boldsymbol{w} = 1$, and set $\boldsymbol{s}_j = \boldsymbol{w}^T\boldsymbol{Y}$.

Greedy: A Greedy Algorithm to Reconstruct $\boldsymbol{X}$ and $\boldsymbol{A}$.

1. **REQUIRE:** $\mathcal{S} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_T\} \subset \mathbb{R}^p$.
2. For $i = 1 \dots n$

   REPEAT

   $l \leftarrow \arg\min_{\boldsymbol{s}_l \in \mathcal{S}} \|\boldsymbol{s}_l\|_0$, breaking ties arbitrarily

   $\boldsymbol{x}_i = \boldsymbol{s}_l$

   $\mathcal{S} = \mathcal{S} \backslash \{\boldsymbol{s}_l\}$

   UNTIL rank$([\boldsymbol{x}_1, \dots, \boldsymbol{x}_i]) = i$
3. Set $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]^T$, and $\boldsymbol{A} = \boldsymbol{Y}\boldsymbol{Y}^T(\boldsymbol{X}\boldsymbol{Y}^T)^{-1}$.

Provable solution in the complete case

$$\text{minimize } \|\boldsymbol{WY}\|_1 \quad \text{subject to} \quad \text{diag}[\boldsymbol{WY}] = \mathbf{1}.$$

If the expected nonzeros per column is smaller than $\sqrt{n}$ the algorithm **succeeds whp**:

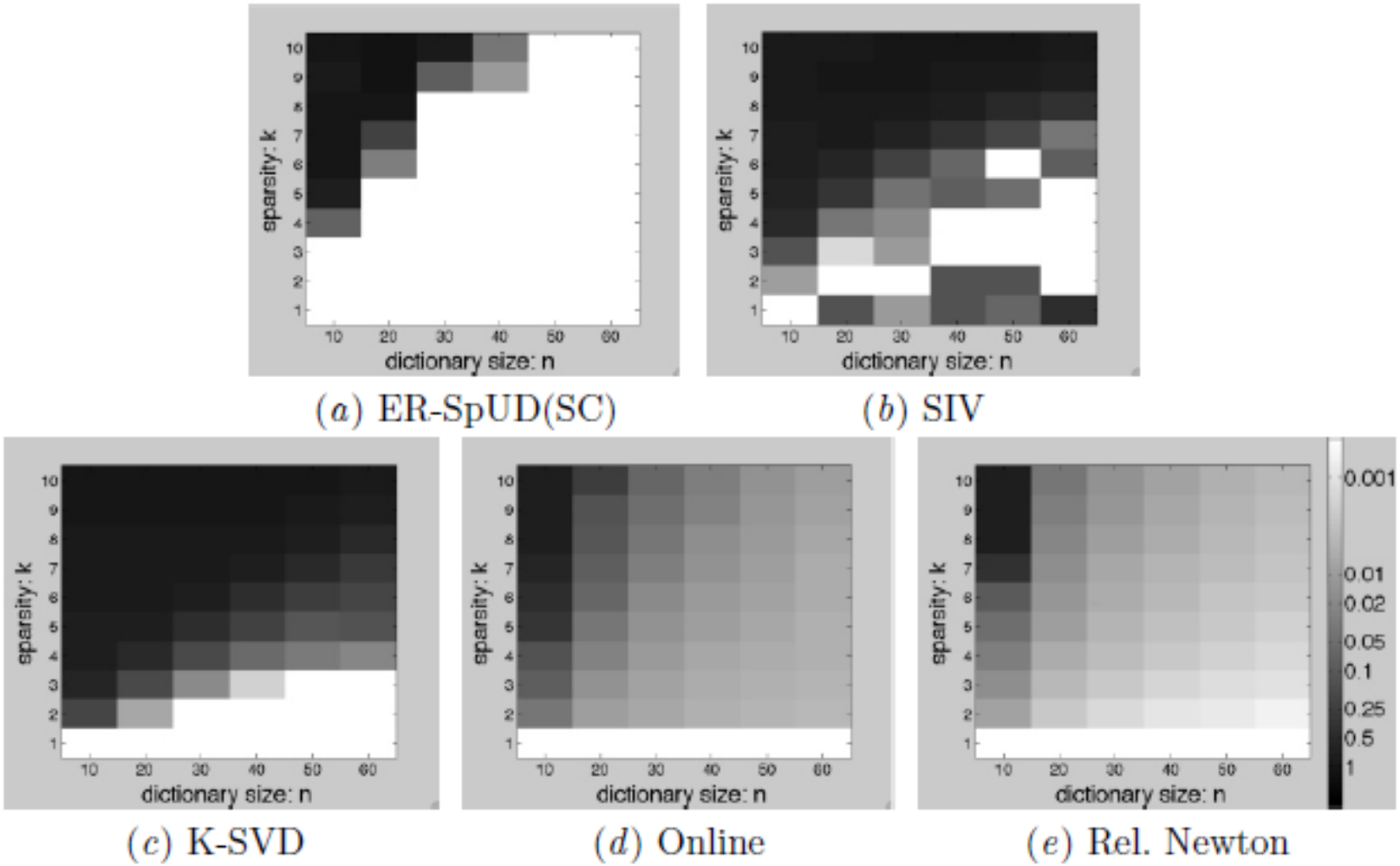
Theorem 12 (Spielman, Wang, W. '12) *(sketch) Let* $\boldsymbol{X}$ Bernoulli(θ)−Rademacher *or* Bernoulli(θ) − Gaussian. *If* $n > n_0$, $p > c_p n^2 \log^2 n$, *and the nonzero probability satisfies*

$$\frac{2}{n} \leq \theta \leq \frac{c}{\sqrt{n}}, \tag{1}$$

with high probability **ER-SpUD (DC)** *recovers all* n *rows of* $\boldsymbol{X}$.

Sample requirement $p > cn^2 \log^2 n$

Does it really work?



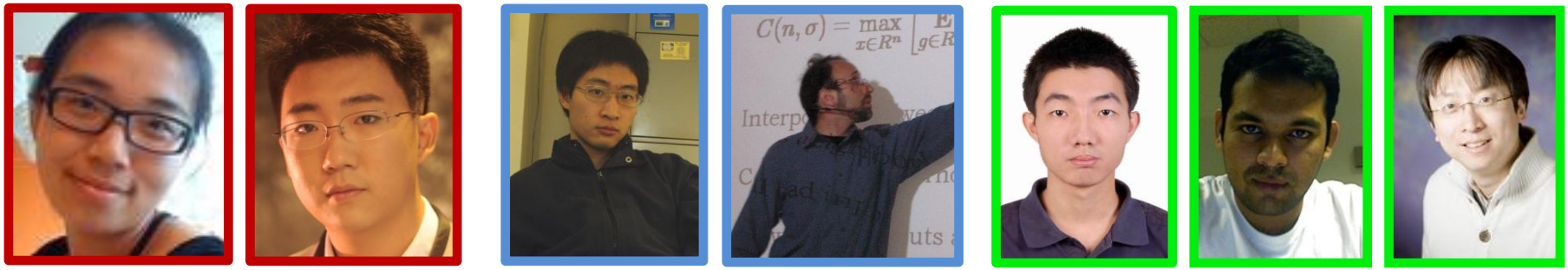
(a) ER-SpUD(SC) (b) SIV

(c) K-SVD (d) Online (e) Rel. Newton

Caveat: exact sparse, noiseless setting.

Provable Representations for Visual Data

Thanks to …



Guaranteed Illumination Models for Nonconvex Objects, Zhang, Mu, Kuo, W., Arxiv '13
Compressive Principal Component Pursuit, W., Ganesh, Min, Ma, I&I '13

Local correctness of ℓ^1-minimization for dictionary learning, Geng, W., Arxiv '11
Exact Recovery of Sparsely-Used Dictionaries, Spielman, Wang, W., COLT '12

Robust Principal Component Analysis? Candes, Li, Ma, W. JACM '11
Robust Face Recognition via Sparse Representation, W., Yang, Ganesh,, Sastry, Ma, PAMI '09
Towards a Practical Automatic Face Recognition … Wagner, W., Ganesh, Zhou, Mobahi, Ma, PAMI '12