Dictionary Learning by ℓ^1-Minimization

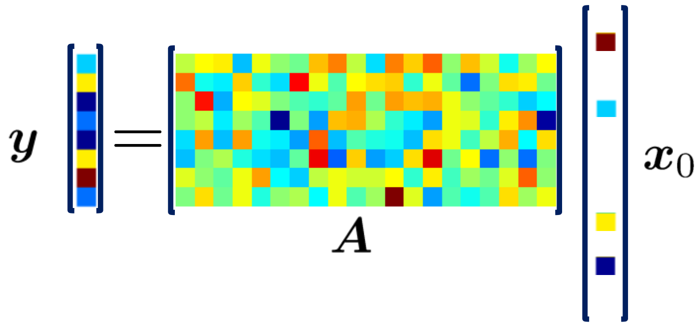
John Wright

Electrical Engineering

Columbia University

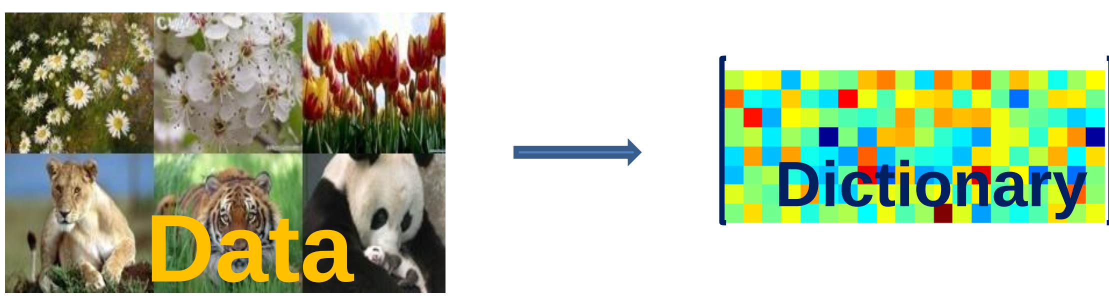
Sparse Approximation

Model:



with $\boldsymbol{x}_0 \in \mathbb{R}^n$ **sparse** - most of the $x_0(i)$ are zero.

Good model for many types of imagery data, especially if we can **learn the dictionary** $\boldsymbol{A} = [\boldsymbol{a}_1 \mid \cdots \mid \boldsymbol{a}_n] \in \mathbb{R}^{m \times n}$:



Motivating Applications (biased sample)

Single image superresolution:

 $= D\textit{ownsample}$ 

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{y}_0$$

reconstruct the original high-resolution image $\boldsymbol{y}_0$:

$$\hat{\boldsymbol{x}} \in \arg\min \|\boldsymbol{x}\|_1 \text{ s.t. } \boldsymbol{y} = \boldsymbol{D}\boldsymbol{A}\boldsymbol{x} \qquad \boldsymbol{y}_0 \approx \boldsymbol{A}\hat{\boldsymbol{x}}$$

Yang, W., Huang and Ma, TIP ‘10

Motivating Applications (biased sample)

High-resolution hyperspectral imaging for cultural heritage:

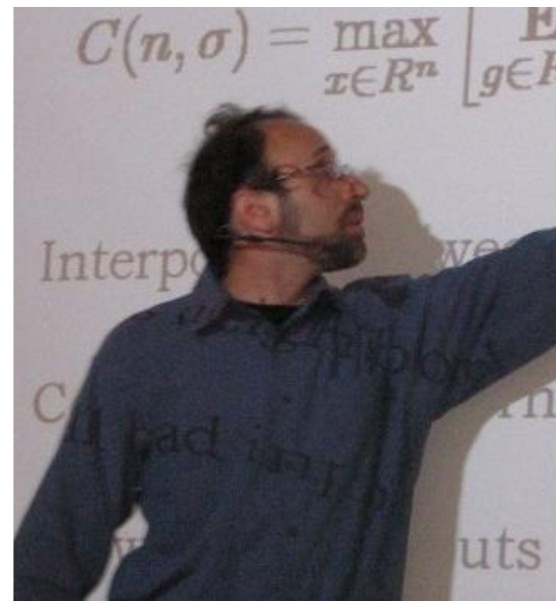
Ultra high-res RGB camera
Moshe Ben-Ezra
Microsoft Research

Buddhist Frescos
Dunhuang, China

Can dictionary learning help overcome hardware limitations?

Kawakami, W., Tai, Ikiuchi, Matsushita, Ben-Ezra, CVPR '11

When do dictionary learning algorithms succeed?



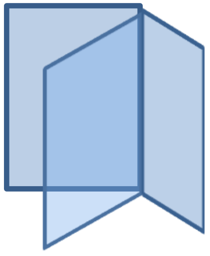
Huan Wang (Yale) **Quan Geng** (UIUC) **Dan Spielman** (Yale)

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 $\boldsymbol{X}$.

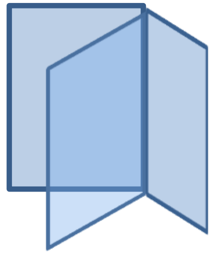
Ambiguities: $(\boldsymbol{A}, \boldsymbol{X})$ or $(\boldsymbol{A}\boldsymbol{\Pi}\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{-1}\boldsymbol{\Pi}^*\boldsymbol{X})$?

Peculiar geometry:

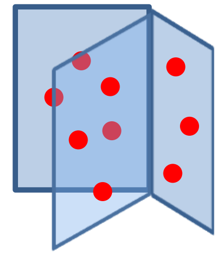


k column subspaces of $\boldsymbol{A}$

When is dictionary learning well-posed?



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

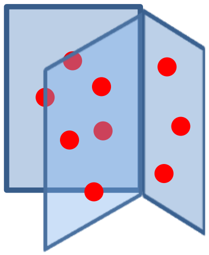


$\boldsymbol{Y}$ k+1 points per subspace

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?



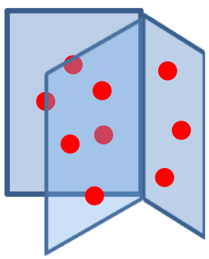
Theorem 2 (Vainsencher, Mannor and Bruckstein '11) *(sketch) If* $\boldsymbol{y} \sim_{iid} \mu$ *on* $\mathbb{S}^{m-1}$, $p > p_0$, $\lambda > \lambda_0$, *then with prob.* $1 - e^{-t}$ *in* $\boldsymbol{Y}$,

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

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

See also [Maurer and Pontil '10].

How can we learn a good dictionary?



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

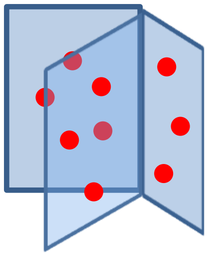
Alternating directions to minimize sparsity surrogate

[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?

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

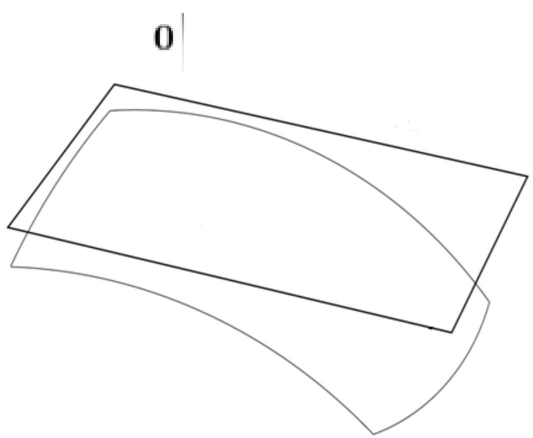
For general $\boldsymbol{A}$, under probabilistic assumptions on $\boldsymbol{X}$, $(\boldsymbol{A}, \boldsymbol{X})$ is a **local minimum whp**:

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 …

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

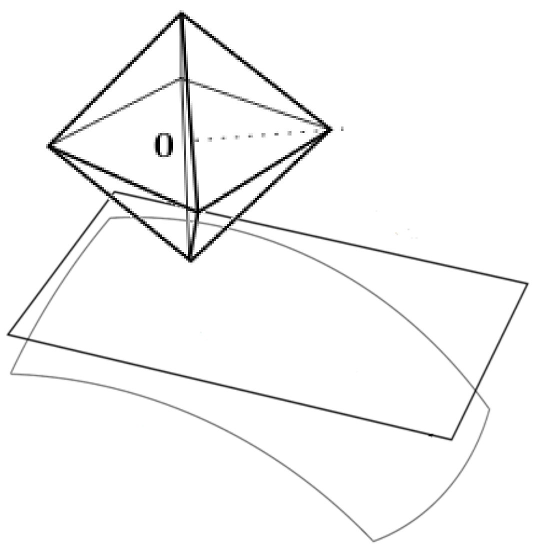


Feasible $(\boldsymbol{A}, \boldsymbol{X})$

Is this obvious?

Maybe … but surprisingly resistant to analysis …

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

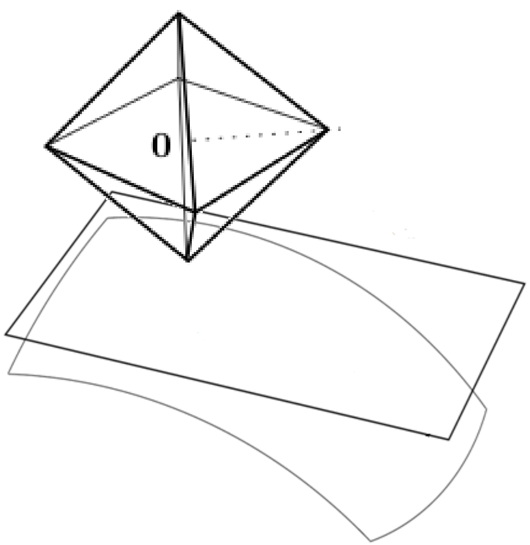


Feasible $(\boldsymbol{A}, \boldsymbol{X})$

Is this obvious?

Maybe … but surprisingly resistant to analysis …

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



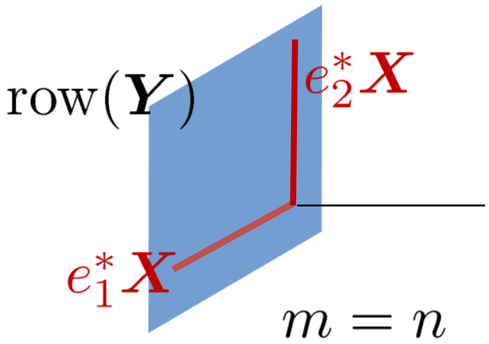
Feasible $(\boldsymbol{A}, \boldsymbol{X})$

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].

Uniqueness – square dictionaries

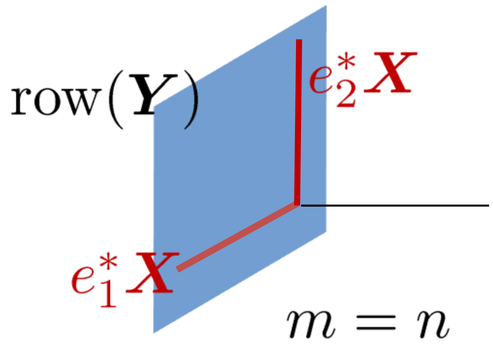


Rows of $\boldsymbol{X}$ are sparse vectors in a known subspace.

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

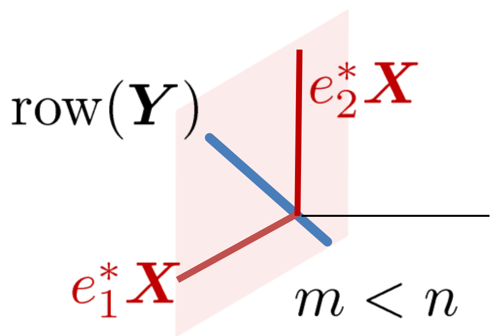
Uniqueness – square dictionaries

Square:



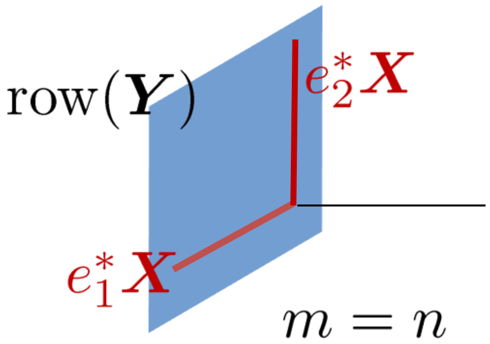
Theorem [Spielman, Wang, W. '11]:
Decomposition essentially unique from $\Omega(n \log n)$ random observations.

Overcomplete:



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

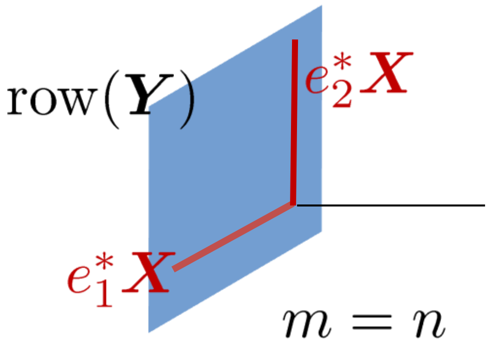
Algorithms – square dictionaries



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

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

Algorithms – square dictionaries



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

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

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

Algorithms – square dictionaries

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

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 $\texttt{rank}([x_1, \dots, x_i]) = i$
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 $\min_{\boldsymbol{w}} \|\boldsymbol{w}^T\boldsymbol{Y}\|_1$ 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 $\boldsymbol{A} = \boldsymbol{Y}\boldsymbol{Y}^T(\boldsymbol{X}\boldsymbol{Y}^T)^{-1}$.

Recovery guarantee – square dictionaries

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

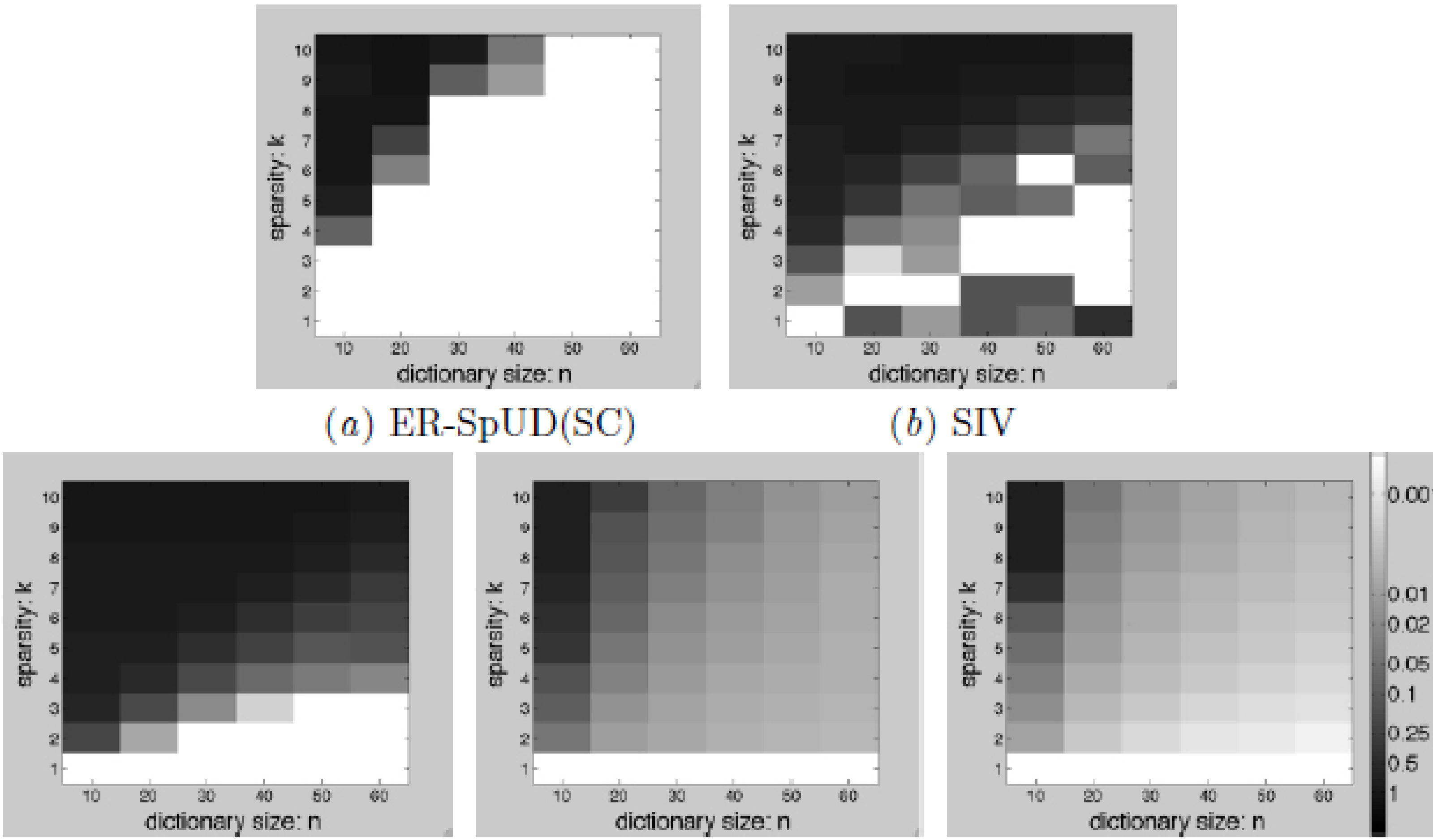
Theorem 5 (Spielman, Wang, W. '12) *(sketch) Let* $\boldsymbol{X}$ Bernoulli(θ)$-$Rademacher *or* Bernoulli(θ) $-$ Gaussian. *If* $n > n_0$, $p > c_p n^2 \log^2 n$, *and the nonzero probability satisfies*

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

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.

Good news / bad news …

If the expected nonzeros per column exceeds $\sqrt{n \log n}$ the algorithm **fails whp**:

Theorem 6 (Spielman, Wang, W. '12) *(sketch) If n large, $p \geq cn$, and the nonzero probability θ satisfies*

$$\theta \geq \sqrt{\frac{\log n}{n}}, \tag{1}$$

then the probability (in $\boldsymbol{X}$) that the algorithm correctly recovers one of the rows is at most n^{-C}.

Theory is almost tight in the sparsity level.

For denser $\boldsymbol{X}$, think about **different constraints.**

Summary and open questions

Two main mathematical results:

Local recovery in the rectangular case

Exact (global) recovery in the square case

Many open questions:

Past the $\sqrt{n}$ barrier?

Noise tolerance, multiple vectors?

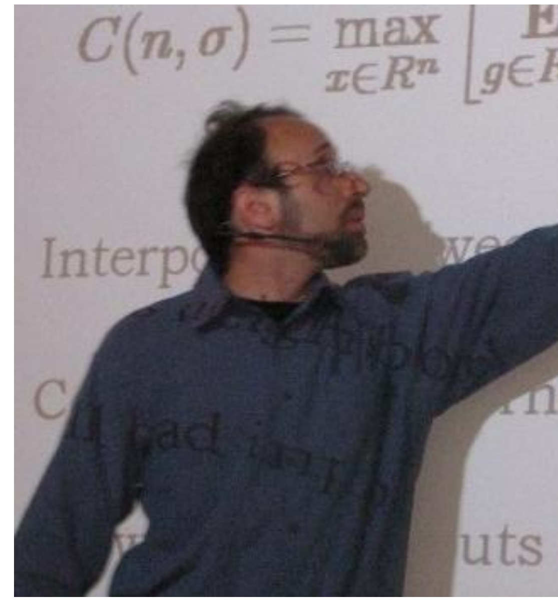
Other coefficient structures?

Dictionary Learning by ℓ^1 Minimization

Thanks to …

Huan Wang (Yale)

Quan Geng (UIUC)



Dan Spielman (Yale)

Local correctness of ℓ^1-minimization for dictionary learning, Geng, W., Arxiv
Exact recovery of sparse dictionaries, Spielman, Wang, W., COLT ’12.