Komprimované snímání a LASSO jako metody zpracování vysocedimenzionálních dat

Jan Vybíral

(Charles University Prague, Czech Republic)

November 2014

VUT Brno

evropský sociální fond v ČR

EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

LASSO

Definition and motivation
Its use in bioinformatics
Its use in material science
Bioinformatics revised

Compressed Sensing

Notions and concepts
Basic results

Connections in EE - further applications

Matrix completion
Separations of features in video
Phase retrieval
MRI

Linear regression

Let $x_1, \ldots, x_N \in \mathbb{R}^d$ be N points in $\mathbb{R}^d$ and $y_1, \ldots, y_N \in \mathbb{R}$

Briefly: $X \in \mathbb{R}^{N\times d}, y \in \mathbb{R}^N$: $\qquad y_i \approx f(x_i)$

Linear regression - least squares (Gauss, Legendre):
$y_i \approx \sum_{j=1}^d \alpha_j X_{ij}$

$$\operatorname*{argmin}_{\alpha\in\mathbb{R}^d} \|y - X\alpha\|_2^2$$

typically, all coordinates of α are non-zero

Regularized linear regression:

$$\operatorname*{argmin}_{\alpha\in\mathbb{R}^d} \|y - X\alpha\|_2 + \lambda\|\alpha\|_2$$

weights between error and size of α

ℓ_1-based methods

Feature selection (Tibshirani, 1996)

LASSO (least absolute shrinkage and selection operator)

$$\underset{\alpha \in \mathbb{R}^d}{\operatorname{argmin}} \; \|y - X\alpha\|_2 + \lambda \|\alpha\|_1, \quad \text{where} \quad \|\alpha\|_1 = \sum_j |\alpha_j|$$

Tends to produce sparse solutions $\alpha \in \mathbb{R}^d$

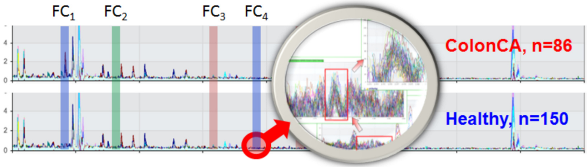
$\lambda > 0$ - regularization parameter

$\lambda \geq \lambda_0$: $\alpha = 0$

$\lambda \to 0$: α goes to least square solution

LASSO in Bioinformatics

- with Tim Conrad, Christoff Schütte (FU Berlin), Gitta Kutyniok (TU Berlin)
- Early diagnosis of a disease - from blood samples!
- Mass Spectrometry - snap shot of proteome

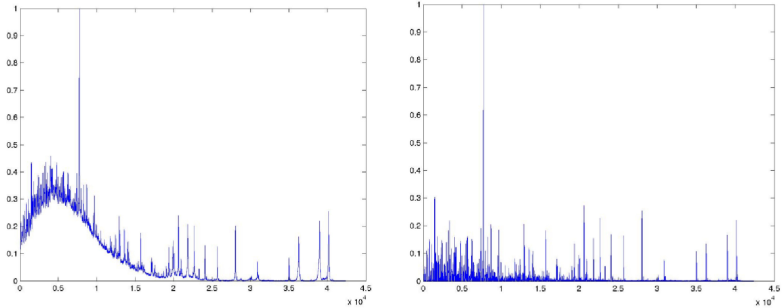


- Very noisy data

$x_1, \dots, x_{100} \in \mathbb{R}^{40000}$ 100 healthy patients

$x_{101}, \dots, x_{200} \in \mathbb{R}^{40000}$ 100 sick patients

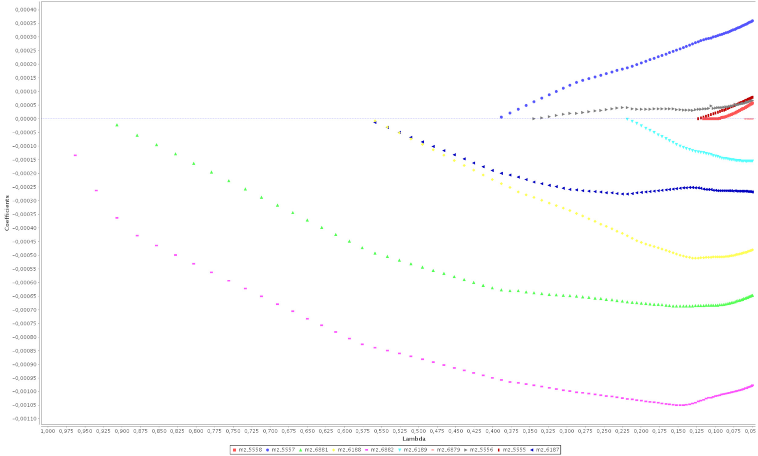
$X \in \mathbb{R}^{200x40000}$, $y_1 = \cdots = y_{100} = 1$ and $y_{101} = \cdots = y_{200} = -1$



(with and without baseline)

- Methods are first tested on synthetic data (with limited amount of artificial and controlled noise)
- Different methods of preprocessing used
- Success rate tested by leave-some-out cross validation
- Rates above 90%, depend on the number of features (ca. 20-50)
- Extensive tests would be necessary (more data points)

Effect of $\lambda > 0$ on the support of ω



LASSO in Material Science

with Luca M. Ghiringhelli, Matthias Scheffler,
Sergey Levchenko (FHI Berlin) and Claudia Draxl (Humboldt U. Berlin)

Classification problem in material science

Task: Given two atoms (i.e. Na & Cl) decide their crystal structure - Zinc blende (ZB) or Rock salt (RS)

Common features: Two atom types form two interpenetrating face-centered cubic lattices

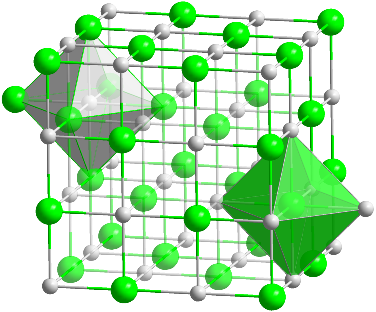
Differences: Relative position of these two lattices. ZB/RS: Each atom's nearest neighbors consist of four/six atoms of the opposite type

Wurtzite: Crystal type very similar to zincblende, materials usually take both the structures depending on conditions

Classification: Given two elements, it is surprisingly hard to predict, which structure they take!

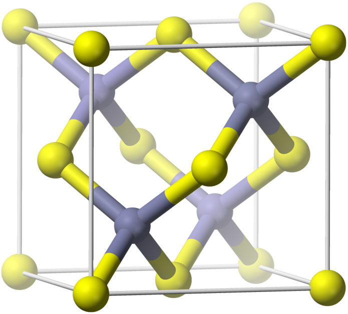
Crystals

NaCl - rocksalt:



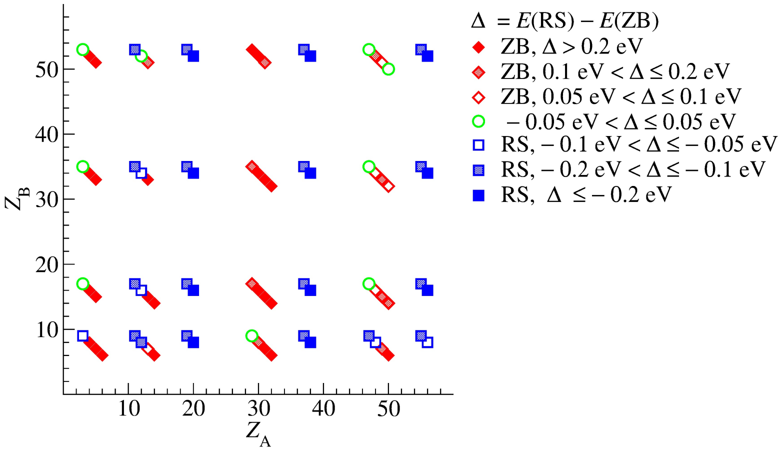
Crystals

ZnS - zinc blende:



Data sets

- 82 compounds of the type AB (NaCl, MgS, AgI, CC, . . .)
- X - 82x2 matrix (columns Z_A, Z_B)
- y - 82x1 vector of $+1$,-1
- $\implies$ classification problem in $\mathbb{R}^2$



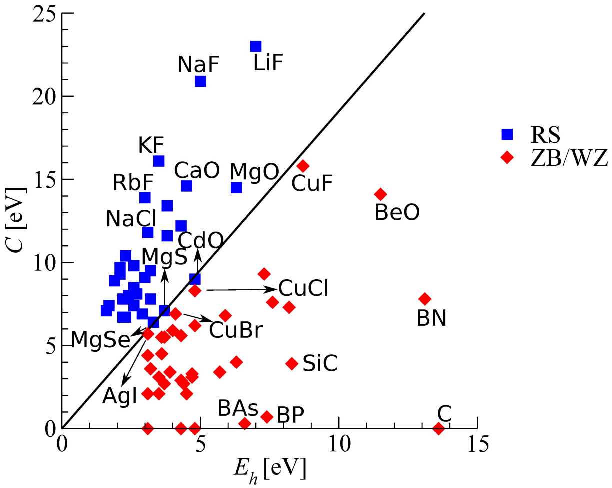
Essentially no machine learning tool can learn such a function from 82 data points only

We replace y by Δ and want to learn $\Delta(AB) = f(Z_A, Z_B)$

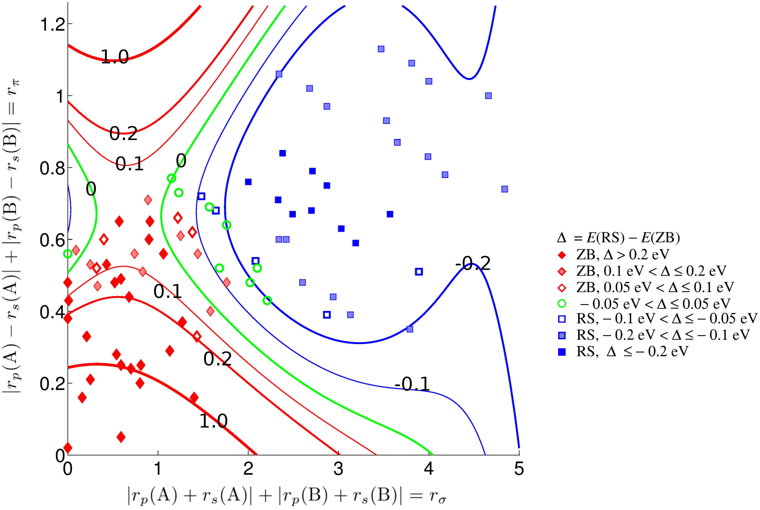
Reduced task - learn Δ from atomic quantities!

- Properties of single atoms
- Easier to calculate
- $r_s(A), r_p(A), r_s(B), r_p(B)$ - orbital radii
- $IP(A), EA(A), IP(B), EA(B)$ - ionization potentials, electroaffinity
- HOMO(A), LUMO(A), HOMO(B), LUMO(B) - energy of Highest Occupied Molecular Orbital and Lowest Unoccupied Molecular Orbital
- . . . *primary features!*

Example 1: Phillips, van Vechten (1969, 1970)



Example 2: Zunger (1980)



Beyond classical data analysis

We construct first physically meaningful quantities:
design of a new, physically motivated kernel

Secondary features - i.e. $1/r_p(A)^2, (r_s(A) - r_p(A))/r_p(B)^3$, etc.

We let LASSO find the best candidates

Due to large coherences $(r_s(A) \approx r_p(A), \dots)$ the selection needs to be stabilized and/or iterated

Results

We found the descriptors

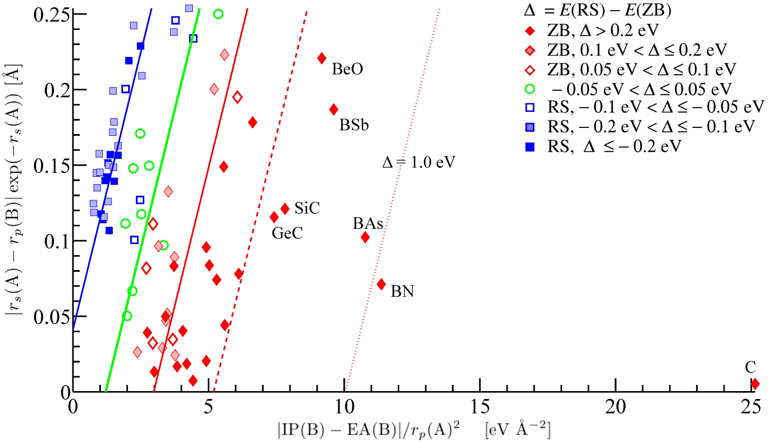
$$\frac{\mathrm{IP(B)} - \mathrm{EA(B)}}{r_p(\mathrm{A})^2}, \frac{|r_s(\mathrm{A}) - r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))}, \frac{|r_p(\mathrm{B}) - r_s(\mathrm{B})|}{\exp(r_d(\mathrm{A}) + r_s(\mathrm{B}))}, \dots$$

Physically reasonable quantities

Goal (for the material science people): Do these descriptors lead to new physics? - Unfortunately, not yet

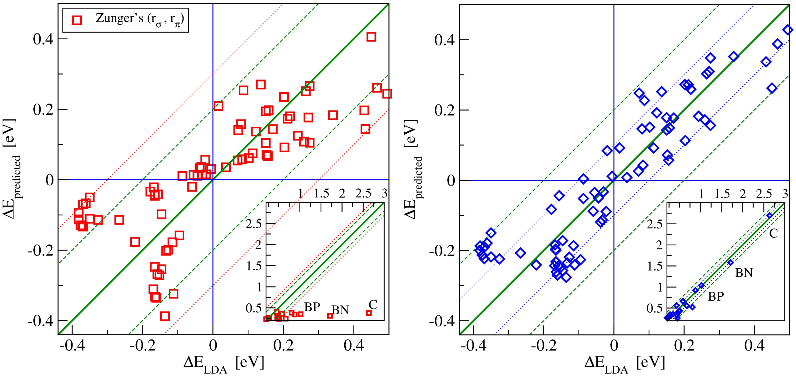
Results

Descriptors found:



Results

Error of a linear fit:



Support Vector Machine

For $\{x_1, \ldots, x_m\} \subset \mathbb{R}^N$ and $\{y_1, \ldots, y_m\} \subset \{-1, 1\}$, the *Support Vector Machine* wants to separate the sets

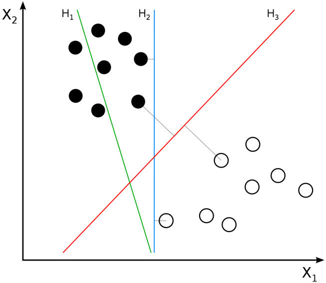
$$\{x_i : y_i = -1\} \quad \text{and} \quad \{x_i : y_i = +1\}$$

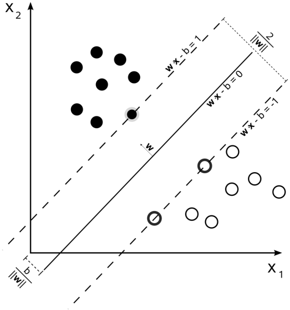
by a linear hyperplane, i.e. finds $w \in \mathbb{R}^N$ and $b \in \mathbb{R}$ with

$$\begin{aligned} \langle w, x_i \rangle - b &> 0 \qquad \text{for} \quad y_i = 1, \\ \langle w, x_i \rangle - b &< 0 \qquad \text{for} \quad y_i = -1. \end{aligned}$$

It maximizes the size of the margin around the separating hyperplane.

Support Vector Machine





$$\min_{w\in\mathbb{R}^N} \sum_{i=1}^{m} (1 - y_i\langle w, x_i\rangle)_+ + \lambda\|w\|_2^2$$

$\lambda > 0$ - a parameter

We want separation based on few coordinates!

1. We want good separation $\implies$ good diagnosis
2. The position of non-zero coordinates should explain the science behind

ℓ_1-SVM replaces $\|w\|_2^2$ by $\|w\|_1$ - promotes the sparsity of w!

Zhu, Rosset, Hastie and Tibshirani (2003)

In bioinformatics: the (few) non-zero components of a sparse w are the "markers" of a disease $\implies$ **causality!?!**

Compressed Sensing

Compressed Sensing

. . . mathematics of LASSO?!?

Sparse recovery

"Simplest" equation in mathematics:

$y = Ax$ for (known) $m \times N$ matrix A and $y \in \mathbb{R}^m$

Task: recover $x \in \mathbb{R}^N$ from y

Studied from many points of view:

Linear algebra: existence, uniqueness

Numerical analysis: stability, speed

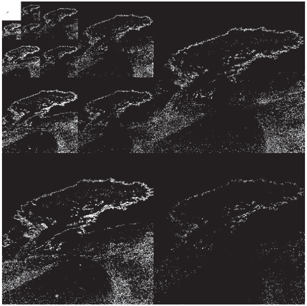
Special methods for structured matrices A

"New" point of view:

. . . we look for a solution x with special structure!

The world is compressible!

Natural images can be sparsely represented by wavelets!. . . JPEG2000



. . . today, we measure all the data (megapixels, i.e. millions), to throw the most of them away!

Setting of Compressed Sensing

Simplified situation:

Let A be an $m \times N$ matrix, and let $x \in \mathbb{R}^N$ be sparse, i.e. with $\|x\|_0 := \#\{i : x_i \neq 0\}$ small. Recover x from $y = Ax$.

Natural assumption:

Given $x \in \mathbb{R}^N$. By experience, we "know" (i.e. expect) that there exists an orthonormal basis Φ with $x = \Phi c$ such that c is sparse

Task:

Let A be an $m \times N$ matrix, let $x = \Phi c \in \mathbb{R}^N$ with Φ an ONB and $\|c\|_0$ small. Recover x from $y = A\Phi c$.

Natural minimization problem: Given an $m \times N$ matrix A and $y \in \mathbb{R}^m$, solve

$$\min_x \|x\|_0 \ \text{ subject to } \ y = Ax$$

This minimization problem is NP-hard!

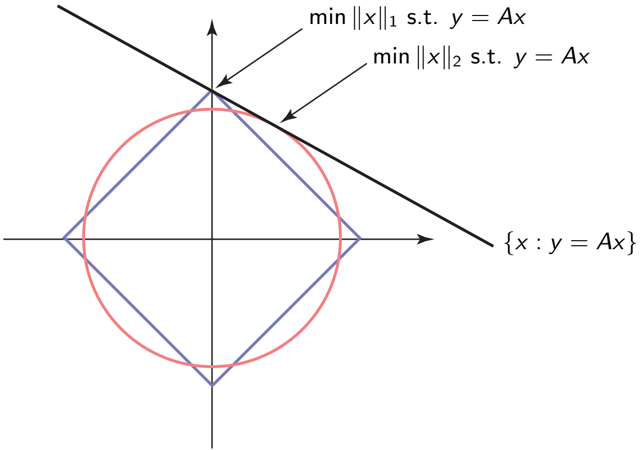
$$\|x\|_p = \Big(\sum_{j=1}^{N} |x_j|^p\Big)^{1/p} : \begin{cases} p \leq 1 - \text{promotes sparsity} \\ p \geq 1 - \text{convex problem} \end{cases}$$

Basis pursuit (ℓ_1-minimization; Chen, Donoho, Saunders - 1998):

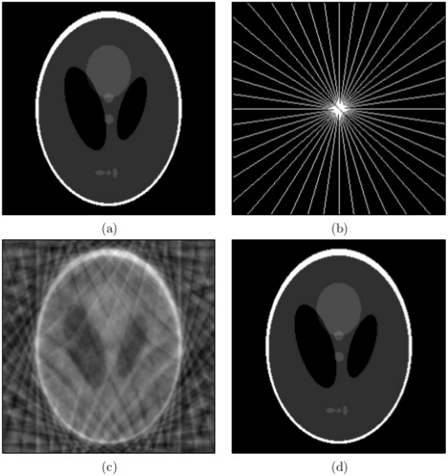
$$\min_x \|x\|_1 \ \text{ subject to } \ y = Ax$$

$\longrightarrow$ This can be solved by linear programming!

ℓ_1 promotes sparsity



red: $x_1^2 + x_2^2 \leq \alpha$ blue: $|x_1| + |x_2| \leq \beta$


(a) (b) (c) (d)

(a) Logan-Shepp phantom, (b) Sampling Fourier coef. along 22 radial lines, (c) ℓ_2 reconstruction, (d) total variation minimization

Source: Candès, Romberg, Tao

Null Space Property

Definition:

$A \in \mathbb{R}^{m \times N}$ has the Null Space Property (NSP) of order s if

$$\|1_\Lambda h\|_1 < \tfrac{1}{2}\|h\|_1 \quad \text{for all } h \in \operatorname{kern}(A) \setminus \{0\} \text{ and for all } \#\Lambda \leq s.$$

Theorem (*Cohen, Dahmen, DeVore - 2008*):

Let $A \in \mathbb{R}^{m \times N}$ and $s \in \mathbb{N}$. TFAE:

(i) Every $x \in \Sigma_s$ is the unique solution of

$$\min_z \|z\|_1 \text{ subject to } Az = y,$$

where $y = Ax$.

(ii) A satisfies the null space property of order s.

Restricted Isometry Property

Definition:
$A \in \mathbb{R}^{m \times N}$ has the Restricted Isometry Property (RIP) of order s with RIP-constant $\delta_s \in (0,1)$ if

$$(1-\delta_s)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta_s)\|x\|_2^2 \qquad \forall x \in \Sigma_s.$$

Theorem (*Cohen, Dahmen, DeVore - 2008; Candès - 2008*):
Let $A \in \mathbb{R}^{m \times N}$ with RIP of order $2s$ with $\delta_{2s} < 1/3$. Then A has NSP of order s.

Directions

Situation:

*Given an $m \times N$ matrix A and an s-sparse $x \in \mathbb{R}^N$,
recover x from $y = Ax$!*

Fundamental (theoretical) questions:

- What is the minimal number $m = m(s, N)$ of measurements?
- For which sensing matrices is the task (uniquely) solvable?
- "Good" algorithms for recovery of x?
- Stability - i.e. "nearly sparse" x's?
- Robustness - i.e. noisy measurements?

Sensing matrices

Random matrices (*Candès, Donoho, et al.; 2006–2011*)
Let A be an $m \times N$-matrix with independent (sub)-gaussian entries. If

$$m \geq C\delta^{-2} s \log(N/s),$$

then A satisfies the RIP of order s with $\delta_s \leq \delta$ with prob. at least

$$1 - 2\exp(-c\delta^2 m) \quad \text{'overwhelmingly high probability'}.$$

Optimality (through high-dimensional geometry):
Stable recovery of s-sparse vectors is possible only for $m \geq Cs\log(N/s)$.

Stability, robustness

The theory can be easily generalized to include

- *stability* (x not sparse but compressible) and
- *robustness* (measurements with noise)

Let $y = Ax + e$, $\|e\|_2 \leq \eta$, where A has the *Robust Null Space Property* of order s. Then

$$x^{\#} := \arg\min_{x} \|x\|_1 \ \text{ subject to } \ \|Ax - y\|_2 \leq \eta$$

satisfies

$$\|x - x^{\#}\|_1 \leq C\sigma_s(x)_1 + D\sqrt{s}\,\eta$$

and

$$\|x - x^{\#}\|_2 \leq \frac{C}{\sqrt{s}}\sigma_s(x)_1 + D\eta.$$

“Matrix completion”, or low-rank matrix recovery

The theory applies to other sorts of sparsity!
x sparse means, that some (unknown) of its possible degrees of freedom are not used (i.e. equal to zero)

The same is true for low-rank matrices!

E. Candès and T. Tao. The power of convex relaxation: near-optimal matrix completion, IEEE Trans. Inform. Theory, 56(5), pp. 2053 - 2080 (2010)

E. Candès and B. Recht. Exact matrix completion via convex optimization, Found. of Comp. Math., 9 (6). pp. 717-772 (2009)

D. Gross, Recovering low-rank matrices from few coefficients in any basis, IEEE Trans. Inform. Theory 57(3), pp. 1548-1566 (2011)

Low-rank matrix recovery

Let $X \in \mathbb{C}^{n_1 \times n_2}$ be a matrix of rank at most r.
Let $y = \mathcal{A}(X) \in \mathbb{C}^m$ be the (linear) measurements of X.

We "want" to solve

$$\operatorname*{arg\,min}_{Z \in \mathbb{C}^{n_1 \times n_2}} \operatorname{rank}(Z) \quad s.t.\ \mathcal{A}(Z) = y.$$

$\operatorname{rank}(Z) = \|(\sigma_1(Z), \sigma_2(Z), \dots)\|_0$ gets replaced by
the nuclear norm $\|Z\|_* = \|(\sigma_1(Z), \sigma_2(Z), \dots)\|_1 = \sum_i |\sigma_i(Z)|$.

The convex relaxation is then

$$\operatorname*{arg\,min}_{Z \in \mathbb{C}^{n_1 \times n_2}} \|Z\|_* \quad s.t.\ \mathcal{A}(Z) = y.$$

Separating features in video's

Some videos (security cameras) can be divided into two parts
- background (= "low rank" component)
- movements (= "sparse" component)

The "intuitive" program

$$\arg\min_{L,S}(\mathrm{rank}L + \lambda\|S\|_0), \quad \text{s.t. } L + S = X.$$

gets replaced by a convex program

$$\arg\min_{L,S}(\|L\|_* + \lambda\|S\|_1), \quad \text{s.t. } L + S = X.$$

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis?, Journal of ACM 58(1), 1-37 (2009)

Data from S. Becker (Caltech)

Separating features in video's: Example

Advanced Background Subtraction

First row:

Left: original image
Middle: low-rank (i.e. predictable) component
Right: sparse component

Second row: similar, quantization effects taken into account, i.e. another term with Frobenius norm added.

Phase retrieval

Setting:

Reconstruct the signal x from the magnitude of its discrete Fourier transform $\hat{x}$

General setting:

x given, $b_k = |\langle a_k, x\rangle|^2, k = 1, \ldots, m$ known, recover x!

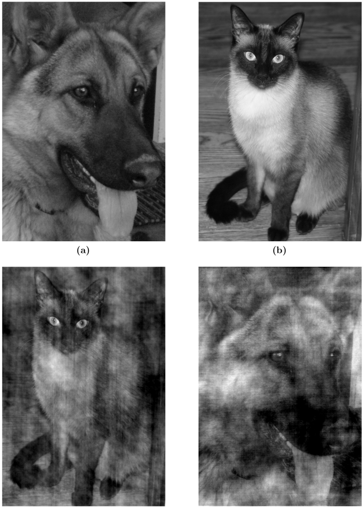
Frequent problem (i.e. astronomy, crystallography, optics), different algorithms exist...

PhaseLift:

quadratic measurements of x are "lifted up" and become linear measurements of the matrix $X := xx^*$:

$$|\langle a_k, x\rangle|^2 = \mathrm{Tr}(x^* a_k a_k^* x) = \mathrm{Tr}(a_k a_k^* xx^*) = \mathrm{Tr}(A_k X) = \langle A_k, X\rangle_F,$$

where $A_k := a_k a_k^*$



(a) (b) (c) (d)

Exchanging Fourier phase while keeping the magnitude

picture: Osherovich

PhaseLift

The "intuitive" problem

$$\begin{array}{ll} \text{find} & X \\ \text{subject to} & (\mathrm{Tr}(A_k X))_{k=1}^m = (b_k)_{k=1}^m \\ & X \geq 0 \\ & \mathrm{rank}(X) = 1 \end{array}$$

gets replaced by a "convex" problem

$$\begin{array}{ll} \text{minimize} & \cancel{\mathrm{rank}(X)} \quad \|X\|_* \\ \text{subject to} & (\mathrm{Tr}(A_k X))_{k=1}^m = (b_k)_{k=1}^m \\ & X \geq 0. \end{array}$$

. . . Matrix recovery problem!

Results

E. Candès, Y. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. SIAM J. on Imaging Sciences 6(1), pp. 199–225, 2011

E. Candès, T. Strohmer and V. Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. Comm. Pure and Appl. Math. 66, pp. 1241–1274, 2011

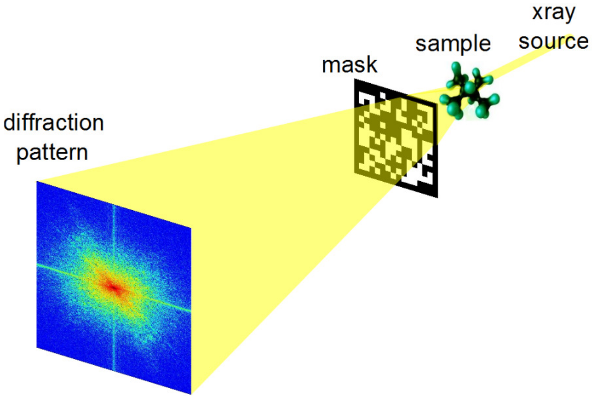
E. Candès and X. Li. Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. To appear in Found. of Comp. Math.

Theorem (Candès, Li, Strohmer, Voroninski, 2011)

If a_k's are chosen independently on the sphere and $m \geq CN$ (not $N \log N$!), then the unique solution of the convex problem is $X = xx^*$ with high probability.

The reconstruction is robust w.r.t. noise!
Version for x sparse!

Implementation of random measurements



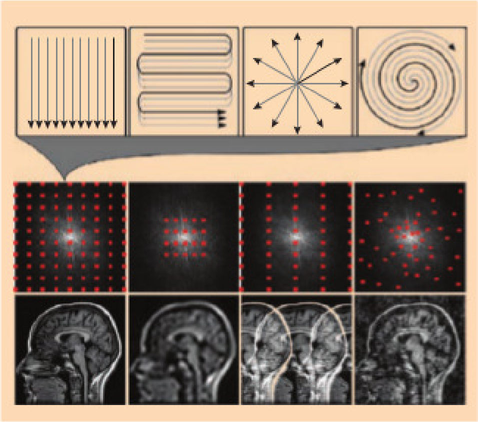
Magnetic Resonance Imaging

MRI exhibits several important features, which suggest using CS:

1. MRI images are naturally sparse (in an appropriate transform and domain).
2. MRI scanners acquire encoded samples, rather than direct pixel samples.
3. Sensing is "expensive" (damage to patient, costs).
4. Processing time does not play much role.

MRI applies additional magnetic fields on top of a strong static magnetic field. The signal measured $s(t)$ is the Fourier transform of the object sampled at certain frequency $\bar{k}(t)$.

How to choose the frequencies, to allow for fast and high-quality recovery?



Different shapes in the k space correspond to sampling of different Fourier coefficients

Literature

- S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, Birkhäuser/Springer 2013
  - Recent, but standard textbook
  - Detailed presentation
- H. Boche, R. Calderbank, G. Kutyniok, and J. V., *A Survey of compressed sensing*, Birkhäuser/Springer, to appear.
  - Short survey
  - 25 pages of basic theory
  - 15 pages of extensions
  - The most important proofs simplified as much as possible
  - Freely available
- Video-lecture of E. Candès from ICM 2014, available on youtube

Thank you for your attention!