Compressed sensing and low-rank optimization

Roman Schutski

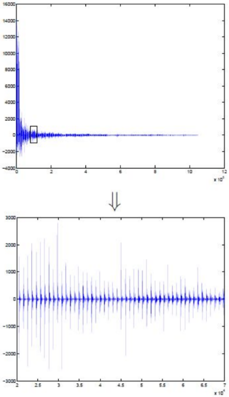
Skoltech

November 27, 2018

Natural signals are sparse

1 megapixel image

wavelet coefficients



zoom in

Implication

We can discard most of small coefficients without loosing perceptual quality

Compressing data

- Calculate 1,000 000 wavelet/cosine/other coefficients of the image
- Drop all but 25,000 largest coefficients
- Invert the transformation

1 megapixel image

25k term approximation

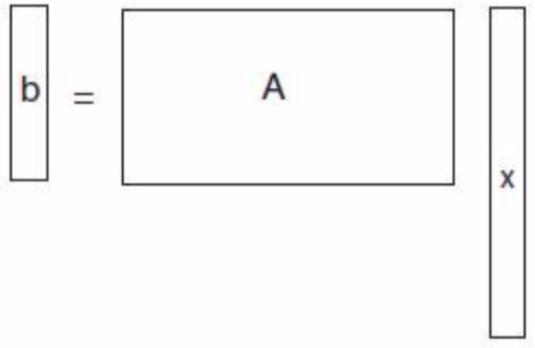
Usual path: collect full $\to$ transform $\to$ shrink

Idea

What if we try to work with a limited subset of data?

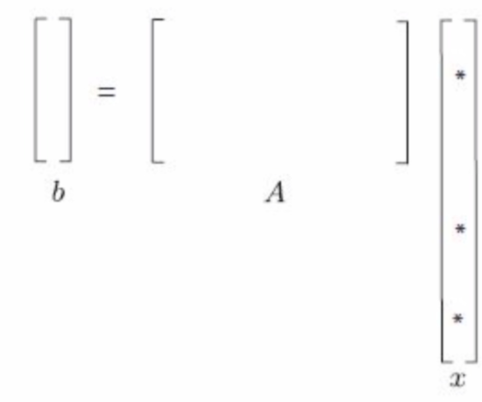
Underdetermined problems

- Have signal $x \in \mathbf{R}^n$, sample a linear combination $A \in \mathbf{R^{m \times n}}$ of entries into $b \in \mathbf{R}^m$, where $m << n$



- In general it is not possible to solve for x by the fundamental theorem of algebra

Special structure



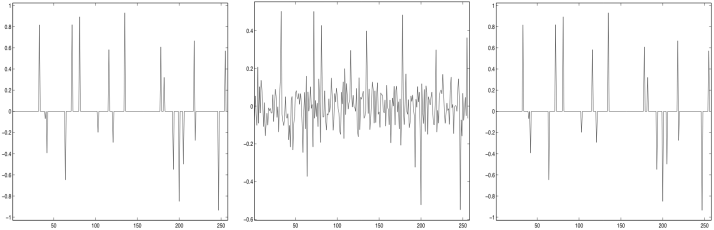
If unknowns are assumed

- Sparse
- Low-rank

this *may* be possible by convex optimization

Example

- Given $n = 256, \ m = 128$
-
```python
A = numpy.randn(m, n)
x = scipy.sparse.random(n, 1, density=0.1)
b = A.dot(x)
```



$$\begin{cases} \min\limits_x \|x\|_0 \\ \text{s.t. } Ax = b \end{cases} \qquad \begin{cases} \min\limits_x \|x\|_2 \\ \text{s.t. } Ax = b \end{cases} \qquad \begin{cases} \min\limits_x \|x\|_1 \\ \text{s.t. } Ax = b \end{cases}$$

(a) exact　　(b) ℓ_2-minimization　　(c) ℓ_1-minimization

Linear programming formulation

L_0 norm

- $||x||_0 =$ number of nonzero elements in x
- Not convex, NP-hard to find a minimum

L_1 norm

- $||x||_1 = \sum_i |x_i|$
- Convex

$$\begin{array}{ll} \text{minimize} & \sum_i |x_i| \\ \text{subject to} & Ax = b \end{array}$$

is equivalent to

$$\begin{array}{ll} \text{minimize} & \sum_i t_i \\ \text{subject to} & Ax = b \\ & -t_i \leq x_i \leq t_i \end{array}$$

with variables $x, t \in \mathbb{R}^n$

$$x^\star \text{solution} \qquad \Longleftrightarrow \qquad (x^\star, t^\star = |x^\star|) \text{ solution}$$

When does this work?

- The field is rapidly evolving and there is no "clean" theory yet
- The requirement $A_{ij} \in \mathcal{N}(0,1)$ is important
- I will highlight only some aspects of the following:
  - How many measurements needed to get solution?
  - What if x is not completely sparse?
  - For what matrices A sparse problem is solvable?
- Not covered:
  - Uniqueness of solutions/properties of L_1-norm relaxation/many other aspects

How many measurements are needed

Result 1 (Candes, Romberg, Tao '06; Donoho, '06)

Suppose $x \in \mathbf{R^n}$ and $||x||_0 = s$ (x is s-sparse). Let $A \in \mathbf{R}^{n \times m}$, $A_{ij} \in \mathcal{N}(0,1) \;\; \forall i, j$, where

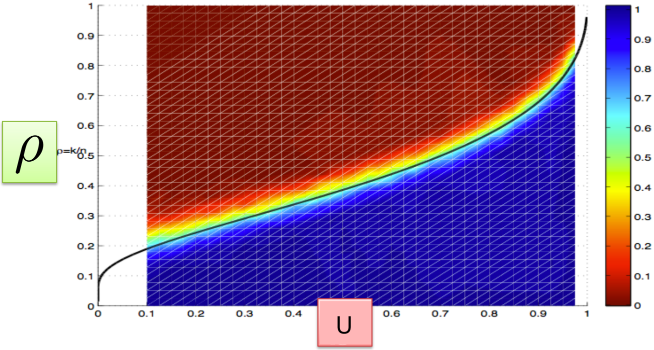
$$m \geq s \cdot log \frac{n}{s}$$

With high probability over the choice of A the solution of $Ax = b$ can be recovered.

Let us define $\frac{m}{n} = u$ (undersampling fraction) and $\frac{s}{n} = \rho$ (sparsity fraction).

Phase transitions

- Under the curve is almost complete recovery, over the curve is almost complete failure!



Building phase transition curve

For a particular matrix A and reconstruction algorithm:

- Choose ρ, u
- Monte Carlo simulation of many problem instances
- Count ratio of successes vs. total problem instances

What if x is not exactly sparse?

Result 2 (C. Romberg and Tao)

Suppose we have a setup of the previous theorem, x is not exactly sparse. If

$$m \geq s \cdot log \frac{n}{s}$$

and $\hat{x}$ is a s-sparse solution, then

$$||\hat{x} - x||_2 \leq ||x - x_s||_1$$

where x_s contains s largest values of x

Corollary

We can look for sparse solutions when x contains noise or is only approximately sparse!

Corollary: Sparse approximation for noisy data

- x is s-sparse, inaccurate measurements: z error term (stochastic or deterministic)

$$b = Ax + z, \text{ with } \|z\|_{\ell_2} \leq \epsilon$$

- Recovery via the LASSO:

$$\min \|\hat{x}\|_{\ell_1} \text{ s. t. } \|A\hat{x} - b\|_{\ell_2} \leq \epsilon$$

- By the Result 2, the error of the final solution is:

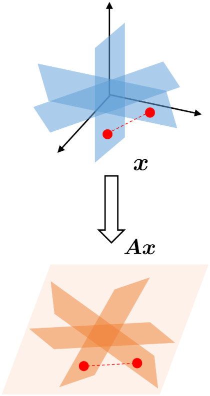
$$\|\hat{x} - x\|_{\ell_2} \lesssim \frac{\|x - x_s\|_{\ell_1}}{\sqrt{s}} + \epsilon = approx.error + measurement\ error$$

What sampling matrices A allow sparse solutions?

- Restricted Isometry Property: suppose x is s-sparse. Then, the smallest possible ϵ_{2s}, s.t.

  $(1-\epsilon)||x_1 - x_2||_F \leq ||Ax_1 - Ax_2||_2^2 \leq (1+\epsilon)||x_1 - x_2||_F$ for all x_1, x_2 is the 2s-isometry constant of A.

- With a picture equivalently: $(1-\epsilon) \leq \frac{||Ax_1 - Ax_2||_2^2}{||x_1 - x_2||_F} \leq (1+\epsilon)$



Restricted Isometry Property

- RIP property is a sufficient condition for sparse algorithms convergence/solution uniqueness
- If $\epsilon_{2s} < 1$ then L_0 minimization problem has a unique sparse solution
- if $\epsilon_{2s} < (\sqrt{2} - 1) = 0.414$ then L_1 relaxed problem has a **the same** unique solution.
- Examples of "good" matrices A:
  - Random $m \times n$ matrices with iid elements if $m >> s\dot{l}og(n/s)$
  - Random $m \times n$ partial DFT matrices if $m >> s\dot{l}og^4(n)$
  - etc

Extending sparsity concept

» **Sparse Vectors:** linear combination of standard basis

$$\mathbf{x} = \sum_i c_i \mathbf{e_i}$$

» **Low Rank Matrices:** linear combination of rank-1 matrices

$$\mathbf{X} = \sum_i c_i \mathbf{u_i} \mathbf{v_i}^T$$

- **From sparsity to low rank matrices:** the problem is still about finding a sparse solution, but in terms of singular values

$$X \longrightarrow \sigma(X)$$

Netflix problem

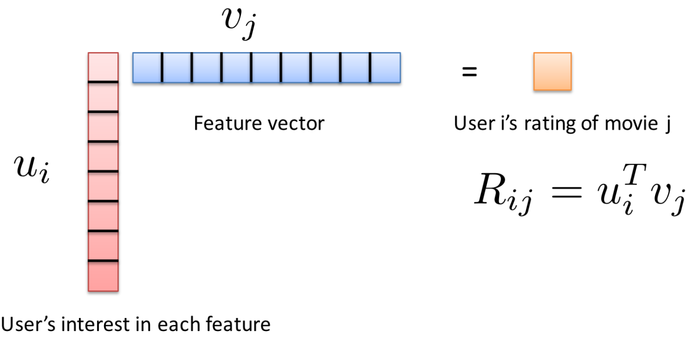
- Predict ratings of movies across all users

$$R \in \mathbb{R}^{m \times n} \qquad \begin{array}{cc} \text{Movies} & \\ \begin{bmatrix} 2 & 3 & ? & ? & 5 & ? \\ 1 & ? & ? & 4 & ? & 3 \\ ? & ? & 3 & 2 & ? & 5 \\ 4 & ? & 3 & ? & 2 & 4 \end{bmatrix} & \text{Users} \end{array}$$

- Only know R_{ij} for $i, j \in \Omega$
- Don't have ratings of every movie from every user

Low rank matrix decomposition

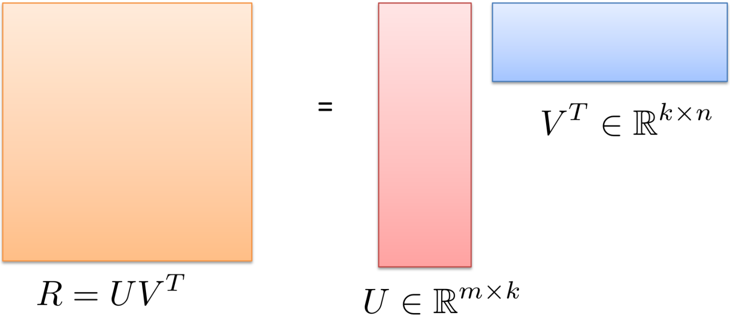
- Can "explain" a movie rating by a small (k) number of features
  - Actors, genre, storyline, length, year, ...
- Each user has a preference for the features



$$R_{ij} = u_i^T v_j$$

Low rank matrix decomposition

- Can "explain" a movie rating by a small (k) number of features
  - Actors, genre, storyline, length, year, ...
- Each user has a preference for the features
- Matrix R is low rank, with rank $k << m$, $k << n$



Low rank matrix decomposition

- Given that the true R is low rank, find a matrix X that is low rank and agrees with R at the observed entries:

$$\begin{array}{ll} \underset{X}{\text{minimize}} & \text{rank } X \\ \text{subject to} & \mathcal{A}(X) = y \\ & (X_{ij} = R_{ij} \quad \forall (i,j) \in \Omega) \end{array}$$

- Rank(X) is not a convex function!

Low rank matrix decomposition

- Given that the true R is low rank, find a matrix X that is low rank and agrees with R at the observed entries:

$$\begin{array}{ll} \underset{X}{\text{minimize}} & \text{rank } X \\ \text{subject to} & \mathcal{A}(X) = y \end{array}$$

$$(X_{ij} = R_{ij} \quad \forall (i,j) \in \Omega)$$

- Nuclear norm instead of the rank function (Recht, Fazel, Parrilo):

$$\boxed{||X||_* = \sum_i \sigma_i}$$

(sum of singular values)

Low rank algorithm intuition

- How can we solve the low rank matrix completion problem?
- **Intuition:**
  - A low rank matrix has a small number of non-zero singular values
  - We see a linear mixture of these singular values (through SVD)
  - Apply soft-thresholding iteratively on the singular values of X
- **Projection onto convex sets:**
  - Take the SVD: $X = P\Sigma Q^T$ - not low rank
  - Soft Threshold: $\hat{\Sigma} = S_\lambda(\Sigma)$
  - Form new matrix: $\hat{X} = P\hat{\Sigma}Q^T$ - low rank but inconsistent with R_{ij}
  - Enforce constraints (replace entries): $\hat{X}_{ij} = R_{ij}$

When does this work

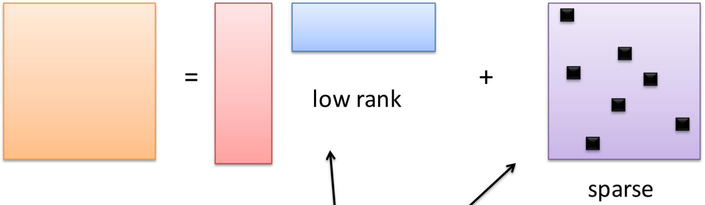
- When the rows/columns of R are incoherent. More specifically, if $\forall X, \quad rank(X) \leq r, \quad \exists \epsilon < 0.414$ s.t.:

$$\boxed{(1-\epsilon)||X||_F^2 \leq ||\mathcal{A}(X)||_2^2 \leq (1+\epsilon)||X||_F^2}$$

  The latter condition is Restricted Isometry Property, satisfied if entries of R are at random positions
- **Theorem**:
  - If X is rank-r and RIP holds for $\mathcal{A}$ and $Y = \mathcal{A}(X)$
  - Then "Projection onto convex sets" converges to the optimum of the problem
- **Uniqueness**:
  - There are multiple solutions to the low-rank problem, but only one is sparse

Low-rank + sparse decomposition

- Low-rank matrix with sparse errors

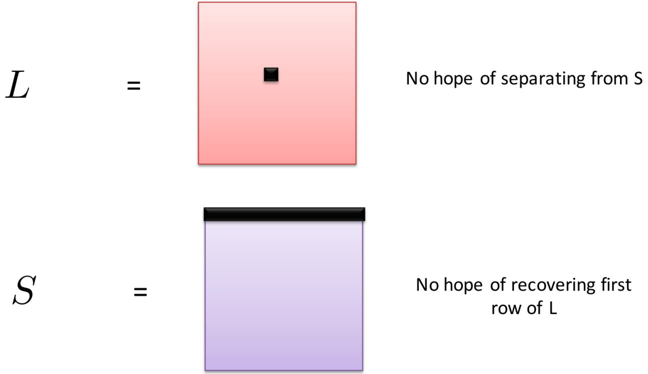


$$R = L + S$$

- Given R, find L and S exactly
- Not a well-posed problem in general

Low-rank + sparse decomposition

- Bad cases:
  - L is low-rank **and** sparse
  - rows/columns of S are coherent



- Exact conditions are given in Candès E., Li X., Ma Y., Wright J. *Robust principal component analysis?*

Low-rank + sparse decomposition

- Under the conditions presented in (Candès, Li, Ma, Wright) the following convex problem has a unique solution:

$$\begin{aligned} \underset{L,S}{\text{minimize}} \quad & ||L||_* + \lambda ||S||_1 \\ \text{subject to} \quad & R = L + S \end{aligned}$$

Low-rank + sparse decomposition

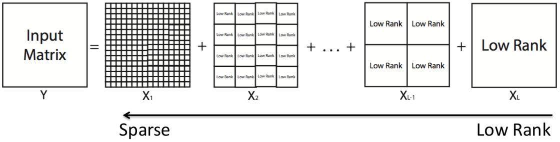
- Application: background separation
  - Background is slowly changing between frames → low rank
  - Fast changing components are rare → sparse

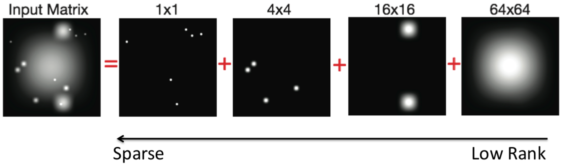
Input Video

Low Rank
+
Sparse

Multiscale low-rank + sparse

- Sum of matrices with increasing scales of correlation (aka wavelets)
- Do decomposition at all scales





- Full details in: F. Ong, M Lustig, *Beyond low rank+ sparse: Multiscale low rank matrix decomposition*, 2016

Multiscale low-rank + sparse

- Algorithm idea:

$$\begin{aligned} \underset{X_i}{\text{minimize}} \quad & \sum_{i=0}^{L-1} \lambda_i \|X_i\|_{(i)} \\ \text{subject to} \quad & Y = \sum_{i=0}^{L-1} X_i \end{aligned}$$

- "Projection onto convex sets" for blocks:
  - Enforce block low rank for each X_i (Block-wise SVD + iterative soft thresholding)
  - Enforce data consistency

Example of multiscale low-rank + sparse

Input Video

Multi-scale Low Rank