Large-Scale L1-Related Minimization in Compressive Sensing and Beyond

Yin Zhang

Department of Computational and Applied Mathematics
Rice University, Houston, Texas, U.S.A.

Arizona State University

March 5th, 2008

RICE

Outline

Outline:

- CS: Application and Theory
- Computational Challenges and Existing Algorithms
- Fixed-Point Continuation: theory to algorithm
- Exploit Structures in TV-Regularization

Acknowledgments: (NSF DMS-0442065)

- Collaborators: Elaine Hale, Wotao Yin
- Students: Yilun Wang, Junfeng Yang

RICE

Compressive Sensing Fundamental

Recover sparse signal from incomplete data

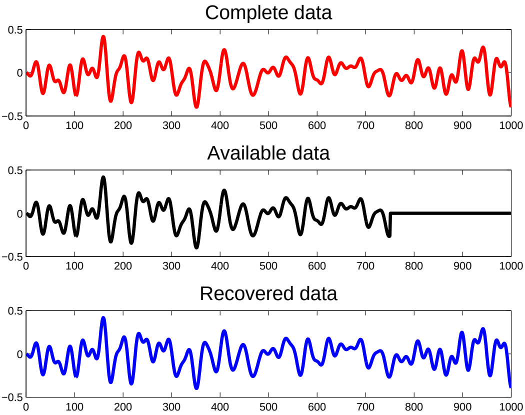
- Unknown signal $x^* \in \mathbb{R}^n$
- Measurements: $Ax^* \in \mathbb{R}^m$, $m < n$
- x^* is sparse (#nonzeros $\|x^*\|_0 < m$)

Unique $x^* = \arg\min\{\|x\|_1 : Ax = Ax^*\}$ $\Rightarrow$ x^* is recoverable

- $Ax = Ax^*$ under-determined, $\min\|x\|_1$ favors sparse x
- Theory: $\|x^*\|_0 < O(m/\log(n/m))$ $\Rightarrow$ recovery for random A (Donoho *et al*, Candes-Tao *et al* ..., 2005)

RICE

Application: Missing Data Recovery



The signal was synthesized by a few Fourier components.

Application: Missing Data Recovery II

Complete data

Available data

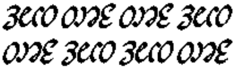
Recovered data

75% of pixels were blacked out (becoming unknown).

RICE

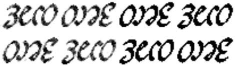
Application: Missing Data Recovery III

Complete data



Available data

Recovered data



85% of pixels were blacked out (becoming unknown).

How are missing data recovered?

Data vector f has a missing part u:

$$f := \begin{bmatrix} b \\ u \end{bmatrix}, \quad b \in \Re^m, \; u \in \Re^{n-m}.$$

Under a basis Φ, f has a representation x^*, $f = \Phi x^*$, or

$$\begin{bmatrix} A \\ B \end{bmatrix} x^* = \begin{bmatrix} b \\ u \end{bmatrix}.$$

Under favorable conditions (x^* is sparse and A is "good"),

$$x^* = \arg\min\{\|x\|_1 : Ax = b\},$$

then we recover missing data $u = Bx^*$.

Sufficient Condition for Recovery

Feasibility: $\mathcal{F} = \{x : Ax = Ax^*\} \equiv \{x^* + v : v \in Null(A)\}$

Define: $S^* = \{i : x_i^* \neq 0\}, \quad Z^* = \{1, \cdots, n\} \setminus S^*$

$$\begin{aligned} \|x\|_1 &= \|x^*\|_1 + (\|v_{Z^*}\|_1 - \|v_{S^*}\|_1) + \\ &\quad (\|x^*_{S^*} + v_{S^*}\|_1 - \|x^*_{S^*}\|_1 + \|v_{S^*}\|_1) \\ &> \|x^*\|_1, \quad \text{if } \|v_{Z^*}\|_1 > \|v_{S^*}\|_1 \end{aligned}$$

x^* is the unique min. if $\|v\|_1 > 2\|v_{S^*}\|_1, \forall v \in Null(A) \setminus \{0\}$.

Since $\|x^*\|_0^{1/2}\|v\|_2 \geq \|v_{S^*}\|_1$, it suffices that

$$\|v\|_1 > 2\|x^*\|_0^{1/2}\|v\|_2, \quad \forall v \in Null(A) \setminus \{0\}$$

RICE

ℓ_1-norm vs. Sparsity

Sufficient Sparsity for Unique Recovery:

$$\sqrt{\|x^*\|_0} < \frac{1}{2}\frac{\|v\|_1}{\|v\|_2}, \quad \forall v \in Null(A) \setminus \{0\}$$

By uniqueness,

$$x \neq x^*, \; Ax = Ax^* \Rightarrow \|x\|_0 > \|x^*\|_0.$$

Hence,

$$\begin{aligned} x^* &= \arg\min\{\|x\|_1 : Ax = Ax^*\} \\ &= \arg\min\{\|x\|_0 : Ax = Ax^*\} \end{aligned}$$

i.e., minimum ℓ_1-norm implies maximum sparsity.

In most subspaces, $\|v\|_1 \gg \|v\|_2$

In $\mathbb{R}^n$, $1 \le \frac{\|v\|_1}{\|v\|_2} \le \sqrt{n}$. However, $\|v\|_1 \gg \|v\|_2$ in most subspaces (due to concentration of measure).

Theorem: (Kashin 77, Garnaev-Gluskin 84)

Let $A \in \mathbb{R}^{m\times n}$ be standard iid Gaussian. With probability above $1 - e^{-c_1(n-m)}$,

$$\frac{\|v\|_1}{\|v\|_2} \ge \frac{c_2\sqrt{m}}{\sqrt{\log(n/m)}}, \quad \forall v \in Null(A) \setminus \{0\}$$

where c_1 and c_2 are absolute constants.

Immediately, for random A and with high probability

$$\|x^*\|_0 < \frac{Cm}{\log(n/m)} \Rightarrow x^* \text{ is recoverable.}$$

Signs help

Theorem:

There exist good measurement matrices $A \in \mathbb{R}^{m\times n}$ so that if

$$x^* \geq 0 \;\text{ and }\; \|x^*\|_0 \leq \lfloor m/2 \rfloor,$$

then

$$x^* = \arg\min\{\|x\|_1 : Ax = Ax^*, x \geq 0\}.$$

In particular, (generalized) Vandermonde matrices (including partial DFT matrices) are good.

("$x^* \geq 0$" can be replaced by "$sign(x^*)$ is known".)

Discussion

Further Results:

- Better estimates on constants (still uncertain)
- Some non-random matrices are good too
  (e.g. partial transforms)

Implications of CS:

- Theoretically, sample size $n \to O(k \log(n/k))$
- Work-load shift: encoder $\to$ decoder
- New paradigm in data acquisition?
- In practice, compression ratio not dramatic, but
  — longer battery life for space devises?
  — shorter scan time for MRI? ...

RICE

Related ℓ_1-minimization Problems

$$\min\{\|x\|_1 : Ax = b\} \quad \text{(noiseless)}$$

$$\min\{\|x\|_1 : \|Ax - b\| \leq \epsilon\} \quad \text{(noisy)}$$

$$\min \mu\|x\|_1 + \|Ax - b\|^2 \quad \text{(unconstrained)}$$

$$\min \mu\|\Phi x\|_1 + \|Ax - b\|^2 \quad (\Phi^{-1} \text{ may not exist})$$

$$\min \mu\|G(x)\|_1 + \|Ax - b\|^2 \quad (G(\cdot) \text{ may be nonlinear})$$

$$\min \mu\|G(x)\|_1 + \nu\|\Phi x\|_1 + \|Ax - b\|^2 \quad \text{(mixed form)}$$

- Φ may represent wavelet or curvelet transform
- $\|G(x)\|_1$ can represent isotropic TV (total variation)
- Objectives are not necessarily strictly convex
- Objectives are non-differentiable

RICE

Algorithmic Challenges

Large-scale, non-smooth optimization problems with dense data that require low storage and fast algorithms.

- 1k $\times$ 1k, 2D-images give over 10^6 variables.
- “Good" matrices are dense (random, transforms...).
- Often (near) real-time processing is required.
- Matrix factorizations are out of question.
- Algorithms must be built on Av and $A^T v$.

RICE

Algorithm Classes (I)

Greedy Algorithms:

- Marching Pursuits (Mallat-Zhang, 1993)
- OMP (Gilbert-Tropp, 2005)
- StOMP (Donoho et al, 2006)
- Chaining Pursuit (Gilbert et al, 2006)
- Cormode-Muthukrishnan (2006)
- HHS Pursuit (Gilbert et al, 2006)

Some require special encoding matrices.

Algorithm Classes (II)

Introducing extra variables, one can convert compressive sensing problems into smooth linear or 2nd-order cone programs; e.g. $\min\{\|x\|_1 : Ax = b\} \Rightarrow$ LP

$$\min\{e^T x_+ - e^T x_- : Ax_+ - Ax_- = b, x_+, x_- \geq 0\}$$

Smooth Optimization Methods:

- Projected Gradient: GPSR (Figueiredo-Nowak-Wright, 07)
- Interior-point algorithm: ℓ_1-LS (Boyd et al 2007) (pre-conditioned CG for linear systems)
- ℓ_1-Magic (Romberg 2006)

Fixed-Point Shrinkage

$$\min \mu\|x\|_1 + f(x) \quad \Longleftrightarrow \quad x = Shrink(x - \tau\nabla f(x), \tau\mu)$$

where $\qquad Shrink(y,t) = \mathrm{sign}(y) \circ \max(|y| - t, 0)$

Fixed-point iterations:

$$x^{k+1} = Shrink(x^k - \tau\nabla f(x^k), \tau\mu)$$

- directly follows from forward-backward operator splitting (a long history in PDE and optimization since 1950's)
- Rediscovered in signal processing by many since 2000's.
- Convergence properties analyzed extensively

RICE

Forward-Backward Operator Splitting

Derivation:

$$
\begin{aligned}
& \min \mu\|x\|_1 + f(x) \\
\Leftrightarrow \quad & 0 \in \mu\partial\|x\|_1 + \nabla f(x) \\
\Leftrightarrow \quad & -\tau\nabla f(x) \in \tau\mu\partial\|x\|_1 \\
\Leftrightarrow \quad & x - \tau\nabla f(x) \in x + \tau\mu\partial\|x\|_1 \\
\Leftrightarrow \quad & (I + \tau\mu\partial\|\cdot\|_1)x \ni x - \tau\nabla f(x) \\
\Leftrightarrow \quad & \{x\} \ni (I + \tau\mu\partial\|\cdot\|_1)^{-1}(x - \tau\nabla f(x)) \\
\Leftrightarrow \quad & x = shrink(x - \tau\nabla f(x), \tau\mu)
\end{aligned}
$$

$$\min \mu\|x\|_1 + f(x) \quad \Longleftrightarrow \quad x = Shrink(x - \tau\nabla f(x), \tau\mu)$$

RICE

New Convergence Results

The following are obtained by E. Hale, W, Yin and YZ, 2007.

- Finite Convergence: for $k = O(1/\tau\mu)$

$$\begin{aligned} x_j^k &= 0, & \text{if } x_j^* &= 0 \\ \operatorname{sign}(x_j^k) &= \operatorname{sign}(x_j^*), & \text{if } x_j^* &\neq 0 \end{aligned}$$

- Rate of convergence depending on “reduced” Hessian:

$$\limsup_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \leq \frac{\kappa(H_{EE}^*) - 1}{\kappa(H_{EE}^*) + 1}$$

where H_{EE}^* is the sub-Hessian corresponding to $x^* \neq 0$.

The bigger μ is, the sparser x^* is, the faster is the convergence.

Fixed-Point Continuation

For each $\mu > 0$,

$$x = Shrink(x - \tau \nabla f(x), \tau\mu) \quad \Longrightarrow \quad x(\mu)$$

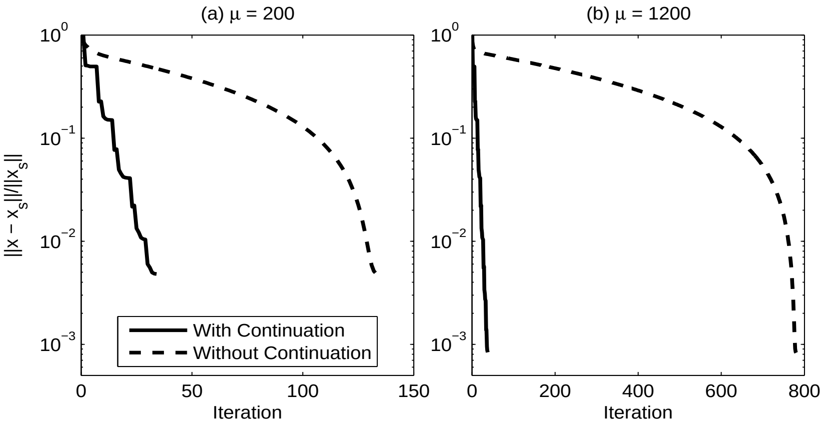
Idea: approximately follow the path $x(\mu)$

FPC:
Set μ to a larger value. Set initial x.
DO until μ it reaches its “right” value

- Adjust stopping criterion
- Start from x, do fixed-point iterations until “stop”
- Decrease μ value

END DO

Continuation Makes It Kick



Discussion

- Continuation make fixed-point shrinkage practical.
- FPC appears more robust than StOMP and GPSR, and is faster most times. ℓ_1-LS is generally slower.
- 1st-order methods slows down on less sparse problems.
- 2-order methods have their own set of problems.
- A comprehensive evaluation is still needed.

RICE

Total Variation Regularization

Discrete (isotropic) TV for a 2D variable:

$$TV(u) = \sum_{i,j} \|(Du)_{ij}\|$$

(1-norm of 2-norms of 1st-order finite difference vectors)

- convex, non-linear, non-differentiable
- suitable for sparse Du, not sparse u

A mixed-norm formulation:

$$\min_u \ \mu TV(u) + \lambda \|\Phi u\|_1 + \|Au - b\|^2$$

Alternating Minimization

Consider linear operator A being a convolution:

$$\min_u \ \mu \sum_{i,j} \|(Du)_{ij}\| + \|Au - b\|^2$$

Introducing $w_{ij} \in \mathbb{R}^2$ and a penalty term:

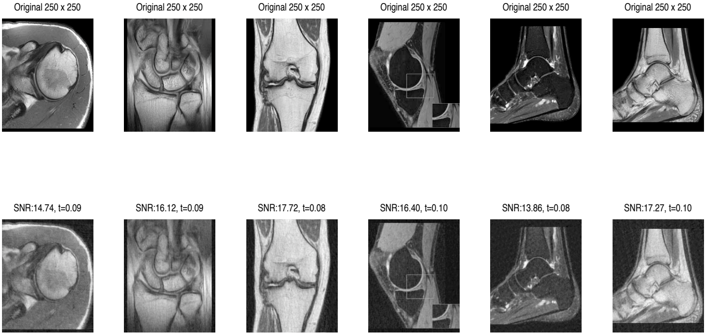
$$\min_{u,w} \ \mu \sum_{i,j} \|w_{ij}\| + \rho\|w - Du\|^2 + \|Au - b\|^2$$

Exploit structure by alternating minimization:

- For fixed u, w has a closed-form solution.
- For fixed w, quadratic can be minimized by 3 FFTs.

(similarly for A being a partial discrete Fourier matrix)

MRI Reconstruction from 15% Fourier Coefficients

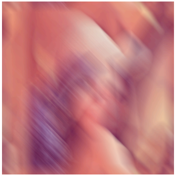


Reconstruction time $\leq$ 0.1s on a Dell PC (3GHz Pentium).

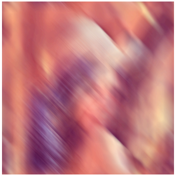
Image Deblurring: Comparison to Matlab Toolbox

Original image: 512x512

Blurry & Noisy SNR: 5.1dB.



deconvlucy: SNR=6.5dB, t=8.9



deconvreg: SNR=10.8dB, t=4.4

deconvwnr: SNR=10.8dB, t=1.4

MxNopt: SNR=16.3dB, t=1.6

512×512 image, CPU time 1.6 seconds

The End

Thank You!