Nonconvex compressive sensing

Rick Chartrand
Los Alamos National Laboratory
Center for Data Driven Modeling and Analysis

joint work with:
Valentina Staneva, Johns Hopkins University
Wotao Yin, Rice University
Kevin Vixie, Washington State University

July 7, 2008
Outline

Motivating example

Compressive sensing

Examples

Algorithms

Local/Global Minimization

Summary
Sparse radiography

Suppose we want to reconstruct an image from samples of its Fourier transform. How many samples do we need?

Sampling along radial lines makes this equivalent to radiographic inversion.

Shepp-Logan phantom

Ω
Motivating example

Nonconvexity

Fewer measurements are needed with nonconvex minimization:

\[
\min_u \| \nabla u \|^p_p, \text{ subject to } \hat{u}\big|_\Omega = \hat{x}\big|_\Omega.
\]
Motivating example

Nonconvexity

Fewer measurements are needed with nonconvex minimization:

$$\min_u \|\nabla u\|_p^p, \text{ subject to } \hat{u}|_\Omega = \hat{x}|_\Omega.$$ 

With $p = 1$, solution is $u = x$ given 18 projections ($|\Omega|/|x| = 6.9\%$).

\[\text{backprojection. 18 views} \quad p = 1, 18 \text{ views}\]
Nonconvexity

Fewer measurements are needed with nonconvex minimization:

$$\min_u \| \nabla u \|^p_p, \text{ subject to } \hat{u}|_\Omega = \hat{x}|_\Omega.$$ 

With $p = 1$, solution is $u = x$ given 18 projections ($\frac{|\Omega|}{|x|} = 6.9\%$).

With $p = 1/2$, 10 projections suffice ($\frac{|\Omega|}{|x|} = 3.8\%$).
Compressive sensing is the reconstruction of sparse (or compressible) signals \( x \) from surprisingly few incoherent measurements \( Ax \):

- We suppose the existence of an operator or dictionary \( \Psi \) such that \( x = \Psi x' \) and most of the components of \( x' \) are (nearly) zero.
- The rows of \( A \) and the columns of \( \Psi \) should not be sparsely expressible in terms of the other.

An undersampled measurement \( Ax \) is tantamount to a compressed version of \( x \). If \( x \) is sufficiently sparse, it can be recovered perfectly.

We exploit the fact that sparsity is mathematically special, yet a general property of natural or human signals.
Compressive sensing is the reconstruction of sparse (or compressible) signals $x$ from surprisingly few incoherent measurements $Ax$:

- We suppose the existence of an operator or dictionary $\Psi$ such that $x = \Psi x'$ and most of the components of $x'$ are (nearly) zero.
- The rows of $A$ and the columns of $\Psi$ should not be sparsely expressible in terms of the other.

An undersampled measurement $Ax$ is tantamount to a compressed version of $x$. If $x$ is sufficiently sparse, it can be recovered perfectly.

- We exploit the fact that sparsity is mathematically special, yet a general property of natural or human signals.
Compressive sensing is the reconstruction of sparse (or compressible) signals $x$ from surprisingly few incoherent measurements $Ax$:

- We suppose the existence of an operator or dictionary $\Psi$ such that $x = \Psi x'$ and most of the components of $x'$ are (nearly) zero.
- The rows of $A$ and the columns of $\Psi$ should not be sparsely expressible in terms of the other.

An undersampled measurement $Ax$ is tantamount to a compressed version of $x$. If $x$ is sufficiently sparse, it can be recovered perfectly.

We exploit the fact that sparsity is mathematically special, yet a general property of natural or human signals.
Compressive sensing is the reconstruction of sparse (or compressible) signals $x$ from surprisingly few incoherent measurements $Ax$:

- We suppose the existence of an operator or dictionary $\Psi$ such that $x = \Psi x'$ and most of the components of $x'$ are (nearly) zero.
- The rows of $A$ and the columns of $\Psi$ should not be sparsely expressible in terms of the other.

An undersampled measurement $Ax$ is tantamount to a compressed version of $x$. If $x$ is sufficiently sparse, it can be recovered perfectly.

- We exploit the fact that sparsity is mathematically special, yet a general property of natural or human signals.
Compressive sensing is the reconstruction of sparse (or compressible) signals $x$ from surprisingly few incoherent measurements $Ax$:

- We suppose the existence of an operator or dictionary $\Psi$ such that $x = \Psi x'$ and most of the components of $x'$ are (nearly) zero.
- The rows of $A$ and the columns of $\Psi$ should not be sparsely expressible in terms of the other.

An undersampled measurement $Ax$ is tantamount to a compressed version of $x$. If $x$ is sufficiently sparse, it can be recovered perfectly.

We exploit the fact that sparsity is mathematically special, yet a general property of natural or human signals.
Let \( x \in \mathbb{R}^N \) be sparse: \( x = \Psi x', \|x'\|_0 = K, K \ll N \).
Optimization for sparse recovery

- Let $x \in \mathbb{R}^N$ be sparse: $x = \Psi x'$, $\|x'\|_0 = K$, $K \ll N$.
- Suppose $\Phi = A\Psi$ is an $M \times N$ matrix, $M \ll N$, with $A$ and $\Psi$ incoherent. For example, $A = (a_{ij})$, i.i.d. $a_{ij} \sim N(0, \sigma^2)$. 

\[
\min_{u} \|u\|_0, \quad \text{s.t.} \quad \Phi u = \Phi x.
\]
\[
\min_{u} \|u\|_1, \quad \text{s.t.} \quad \Phi u = \Phi x.
\]
\[
\min_{u} \|u\|_p, \quad \text{s.t.} \quad \Phi u = \Phi x,
\]

Unique solution is $u = x$ with optimally small $M$, but is NP-hard. $M \geq 2K$ suffices w.h.p.

Can be solved efficiently; requires more measurements for reconstruction.

$M \geq CK \log(\frac{N}{K})$ where $0 < p < 1$. Solvable in practice; requires fewer measurements than $\ell_1$. 

$M \geq C_1(p)K + C_2(p)K \log(\frac{N}{K})$. 

Slide 6 of 17
Let $x \in \mathbb{R}^N$ be sparse: $x = \Psi x'$, $\|x'\|_0 = K$, $K \ll N$.

Suppose $\Phi = A\Psi$ is an $M \times N$ matrix, $M \ll N$, with $A$ and $\Psi$ incoherent. For example, $A = (a_{ij})$, i.i.d. $a_{ij} \sim N(0, \sigma^2)$.

Unique solution is $u = x$ with optimally small $M$, but is NP-hard.

$M \geq 2K$ suffices w.h.p.
Let $x \in \mathbb{R}^N$ be sparse: $x = \Psi x'$, $\|x'\|_0 = K$, $K \ll N$.

Suppose $\Phi = A \Psi$ is an $M \times N$ matrix, $M \ll N$, with $A$ and $\Psi$ incoherent. For example, $A = (a_{ij})$, i.i.d. $a_{ij} \sim N(0, \sigma^2)$.

Unique solution is $u = x$ with optimally small $M$, but is NP-hard.

$M \geq 2K$ suffices w.h.p.

Can be solved efficiently; requires more measurements for reconstruction.

$M \geq CK \log(N/K)$
Let $x \in \mathbb{R}^N$ be sparse: $x = \Psi x'$, $\|x'\|_0 = K$, $K \ll N$.

Suppose $\Phi = A\Psi$ is an $M \times N$ matrix, $M \ll N$, with $A$ and $\Psi$ incoherent. For example, $A = (a_{ij})$, i.i.d. $a_{ij} \sim N(0, \sigma^2)$.

\[
\min_u \|u\|_0, \text{ s.t. } \Phi u = \Phi x.
\]

Unique solution is $u = x$ with optimally small $M$, but is NP-hard. $M \geq 2K$ suffices w.h.p.

\[
\min_u \|u\|_1, \text{ s.t. } \Phi u = \Phi x.
\]

Can be solved efficiently; requires more measurements for reconstruction. $M \geq CK \log(N/K)$ where $0 < p < 1$. Solvable in practice; requires fewer measurements than $\ell^1$.

\[
M \geq C_1(p)K + pC_2(p)K \log(N/K)
\]
The geometry of $\mathcal{L}^p$

$$\min_u \|u\|_p^p, \text{ subject to } \Phi u = \Phi x$$

$p = 2$:

$$\Phi u = \Phi x$$

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.1^p$$
The geometry of $\ell^p$

\[ \min_u \|u\|_p, \text{ subject to } \Phi u = \Phi x \]

\( p = 2: \)

\[ |u_1|^p + |u_2|^p + |u_3|^p = 0.2^p \]
The geometry of $\ell^p$

$$\min_{u} \|u\|_p^p, \text{ subject to } \Phi u = \Phi x$$

$p = 2$:

$$\Phi u = \Phi x$$

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.3^p$$
The geometry of $\ell^p$

\[ \min_u \|u\|_p^p, \text{ subject to } \Phi u = \Phi x \]

$p = 2$:

\[ |u_1|^p + |u_2|^p + |u_3|^p = 0.4^p \]
The geometry of $\ell^p$

$$\min_u \|u\|_p^p, \text{ subject to } \Phi u = \Phi x$$

$p = 2$:

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.5^p$$
The geometry of $\ell^p$

\[ \min_u \|u\|_{\ell^p}^p, \text{ subject to } \Phi u = \Phi x \]

$p = 1$:

\[ \Phi u = \Phi x \]

\[ |u_1|^p + |u_2|^p + |u_3|^p = 0.1^p \]
The geometry of $\mathcal{L}^p$

\[ \min_u \|u\|_p^p, \text{ subject to } \Phi u = \Phi x \]

$p = 1$:

\[ |u_1|^p + |u_2|^p + |u_3|^p = 0.2^p \]
The geometry of $\ell^p$

$$\min_u \|u\|_p^p, \text{ subject to } \Phi u = \Phi x$$

$p = 1$:

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.3^p$$
The geometry of $\mathcal{L}^p$

$$\min_u \|u\|_p^p, \text{ subject to } \Phi u = \Phi x$$

$p = 1$:

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.4^p$$
The geometry of $\ell^p$

$$\min_u \|u\|^p_p, \text{ subject to } \Phi u = \Phi x$$

$p = 1$:

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.5^p$$
The geometry of $\ell^p$

$$\min_u \|u\|^p_p, \text{ subject to } \Phi u = \Phi x$$

$p = 1$:

$$\Phi u = \Phi x$$

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.6^p$$
The geometry of $\ell^p$

$\min_u \|u\|^p_p$, subject to $\Phi u = \Phi x$

$p = 1$:

$|u_1|^p + |u_2|^p + |u_3|^p = 0.7^p$
The geometry of $\ell^p$

$$\min_u \|u\|_p, \text{ subject to } \Phi u = \Phi x$$

$p = 1/2$:

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.1^p$$
The geometry of $\ell^p$}

$$\min_u \|u\|_p^p, \text{ subject to } \Phi u = \Phi x$$

$p = 1/2$:

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.2^p$$
The geometry of $\ell^p$

$$\min_u \|u\|_p, \text{ subject to } \Phi u = \Phi x$$

$p = 1/2$:

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.3^p$$
The geometry of $\ell^p$

$\min_u \|u\|_p$, subject to $\Phi u = \Phi x$

$p = 1/2$:

$\Phi u = \Phi x$

$|u_1|^p + |u_2|^p + |u_3|^p = 0.4^p$
The geometry of $\ell^p$

$p = 1/2$:

$$\min_u \|u\|_p^p, \text{ subject to } \Phi u = \Phi x$$

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.5^p$$
The geometry of $\ell^p$

\[ \min_u ||u||_p, \text{ subject to } \Phi u = \Phi x \]

$p = 1/2$:

\[ |u_1|^p + |u_2|^p + |u_3|^p = 0.6^p \]
The geometry of $\ell^p$

$$\min_u \|u\|_p, \text{ subject to } \Phi u = \Phi x$$

$p = 1/2$:

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.7^p$$
The geometry of $\ell^p$

$$\min_u \|u\|_p^p, \text{ subject to } \Phi u = \Phi x$$

$p = 1/2$:

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.8^p$$
The geometry of $\ell^p$

$$\min_u \|u\|_p, \text{ subject to } \Phi u = \Phi x$$

$p = 1/2$:

$$|u_1|^p + |u_2|^p + |u_3|^p = 0.9^p$$

$$\Phi u = \Phi x$$
The geometry of $\ell^p$

\[ \min_{u} \| u \|_p, \text{ subject to } \Phi u = \Phi x \]

$p = 1/2$:

\[ |u_1|^p + |u_2|^p + |u_3|^p = 1^p \]

\(\Phi u = \Phi x\)
3-D tomography

Non-oriented ellipsoid:

Six radiographic views suffice for exact reconstruction with $p = 1$, three with $p = 1/2$. 
Change 1% of the voxels, randomly. Four views allow an exact reconstruction of the depleted ellipsoid, to identify defects precisely.
A less simple object

Change 1% of the voxels, randomly. Four views allow an exact reconstruction of the depleted ellipsoid, to identify defects precisely.

For objects with piecewise-constant density, far less data is needed than for traditional CT methods.
Cortical activity is reconstructed perfectly from synthetic EEG data, consisting of 256 scalp potential measurements. The synthetic signals have 80 nonzero coefficients from graph-diffusion eigenfunction or wavelet bases. Making the signal only approximately sparse and adding noise results in very little reconstruction error.
We solve the nonconvex optimization problem using an iteratively reweighted $\ell^2$ problem:

$$\min_u \sum_i w_i u_i^2, \quad \text{subject to} \quad \Phi u = b,$$

where the weights are obtained from the previous solution $u^{(n)}$ by

$$w_i = \left( (u_i^{(n)})^2 + \epsilon \right)^{p/2-1}.$$

When $p = 0$, this attempts to minimize

$$\sum_i \log((u_i)^2 + \epsilon).$$
Iteratively reweighted least-squares

We solve the nonconvex optimization problem using an iteratively reweighted $\ell^2$ problem:

$$\min_u \sum_i w_i u_i^2, \text{ subject to } \Phi u = b,$$

where the weights are obtained from the previous solution $u^{(n)}$ by

$$w_i = \left(\left(u_i^{(n)}\right)^2 + \epsilon\right)^{p/2-1}.\text{ When } p = 0, \text{ this attempts to minimize } \sum_i \log\left((u_i)^2 + \epsilon\right).$$

This has an explicit solution:

$$u^{(n+1)} = Q \Phi^T \left(\Phi Q \Phi^T\right)^{-1} b,$$

where $Q$ is the diagonal matrix with entries $1/w_i$. This makes sense with $\epsilon = 0$, with the result being called FOCUSS in this case.
Iteratively reweighted least-squares

We solve the nonconvex optimization problem using an iteratively reweighted $\ell^2$ problem:

$$\min_u \sum_i w_i u_i^2, \quad \text{subject to} \quad \Phi u = b,$$

where the weights are obtained from the previous solution $u^{(n)}$ by

$$w_i = ((u_i^{(n)})^2 + \epsilon)^{p/2-1}. \quad \text{When } p = 0, \text{ this attempts to minimize } \sum_i \log((u_i)^2 + \epsilon).$$

This has an explicit solution:

$$u^{(n+1)} = Q \Phi^T (\Phi Q \Phi^T)^{-1} b,$$

where $Q$ is the diagonal matrix with entries $1/w_i$. This makes sense with $\epsilon = 0$, with the result being called FOCUSS in this case.

Having $\epsilon = 0$ is its downfall.
We let $\epsilon_0 = 1$, then after convergence repeat with $\epsilon_{n+1} = \epsilon_n/10$. This drastically improves the recovery performance.
IRLS with a gradient

Now consider a gradient in the objective. In the unconstrained formulation, this is a lot like regularization in image processing.

\[
\min_u \alpha \|\nabla u\|_p^p + \frac{1}{2} \|\Phi u - b\|_2^2.
\]

IRLS no longer has a useful explicit solution, but the same idea amounts to the lagged diffusivity algorithm of Vogel and Oman, which can be thought of as an approximate Newton method:
Now consider a gradient in the objective. In the unconstrained formulation, this is a lot like regularization in image processing.

$$\min_u \alpha \|\nabla u\|_p^p + \frac{1}{2} \|\Phi u - b\|_2^2.$$  

IRLS no longer has a useful explicit solution, but the same idea amounts to the lagged diffusivity algorithm of Vogel and Oman, which can be thought of as an approximate Newton method: Iterate

$$u_{n+1} = u_n - H_n^{-1}g_n,$$

where

$$H_n = \Phi^T \Phi + \alpha D^T Q^{-1} D$$

is an approximate Hessian, and

$$g_n = H_n u_n - \Phi^T b$$

is the gradient of the functional.

We can solve the equality-constrained problem by simultaneously marching $\epsilon$ and $\alpha$ to zero. For noisy data, we can soften the constraint by leaving $\alpha$ positive.
Why might global minimization be possible?

Consider the $\epsilon$-regularized, constraint-eliminated objective:

$$\displaystyle F_\epsilon(t) = \sum_{i=1}^{N} \left\{ \left[ x_i + (V_t)_i \right]^2 + \epsilon \right\}^{p/2},$$

where $\mathcal{R}(V) = \mathcal{N}(\Phi)$. A moderate $\epsilon$ fills in the local minima.
Consider the $\epsilon$-regularized, constraint-eliminated objective:

$$F_\epsilon(t) = \sum_{i=1}^{N} \left\{ \left[ x_i + (Vt)_i \right]^2 + \epsilon \right\}^{p/2},$$

where $\mathcal{R}(V) = \mathcal{N}(\Phi)$. A moderate $\epsilon$ fills in the local minima.
Why might global minimization be possible?

Consider the $\epsilon$-regularized, constraint-eliminated objective:

$$F_{\epsilon}(t) = \sum_{i=1}^{N} \left\{ \left[ x_i + (Vt)_i \right]^2 + \epsilon \right\}^{p/2},$$

where $\mathcal{R}(V) = \mathcal{N}(\Phi)$. A moderate $\epsilon$ fills in the local minima.
Why might global minimization be possible?

Consider the $\epsilon$-regularized, constraint-eliminated objective:

$$F_\epsilon(t) = \sum_{i=1}^{N} \left\{ [x_i + (V t)_i]^2 + \epsilon \right\}^{p/2},$$

where $\mathcal{R}(V) = \mathcal{N}(\Phi)$. A moderate $\epsilon$ fills in the local minima.
Why might global minimization be possible?

Consider the $\epsilon$-regularized, constraint-eliminated objective:

$$F_\epsilon(t) = \sum_{i=1}^{N} \left\{ \left[ x_i + (Vt)_i \right]^2 + \epsilon \right\}^{p/2},$$

where $\mathcal{R}(V) = \mathcal{N}(\Phi)$. A moderate $\epsilon$ fills in the local minima.

$\epsilon = 0.01$
Why might global minimization be possible?

Consider the $\epsilon$-regularized, constraint-eliminated objective:

$$F_\epsilon(t) = \sum_{i=1}^{N} \left\{ \left[ x_i + (Vt)_i \right]^2 + \epsilon \right\}^{p/2},$$

where $R(V) = \mathcal{N}(\Phi)$. A moderate $\epsilon$ fills in the local minima.

$\epsilon = 0.001$
Compressive sensing allows sparse signals to be recovered with very few measurements.

Nonconvex compressive sensing needs even fewer measurements.

Decreasing $p$ also improves robustness to noise, and speeds up convergence.

Regularizing the objective appears to keep algorithms from converging to nonglobal minima.

Applications: medical imaging tomography, network state inference, streaming data reduction, ...

math.lanl.gov/~rick