Extensions to Sparsity Promotion

Xiang Li and Felix J. Herrmann
Dimensionality Reduction for Sparsity Promotion

Tim Lin and Felix J. Herrmann
Simultaneous Estimation of Green's function and Source Wavelet

Aleksandr Y. Aravkin, James V. Burke, and Michael P. Friedlander
Robust and Sparsity Promoting Formulations

SINBAD, 6 December 2011

Outline

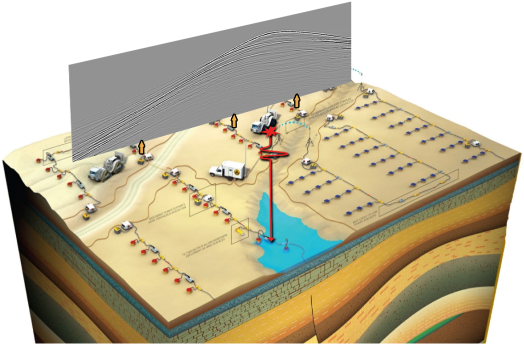
- Sparsity promotion in geophysics via SPGℓ_1
  - a few examples of sparsity promotion in geophysics
  - exploiting Pareto tradeoff curve and inexact Newton method
- Dimensionality Reduction for Sparsity Promotion
  - simultaneous and sequential subsampling
  - modifying SPGℓ_1 to take advantage of dimensionality reduction
- Simultaneous Estimation of Green's Function and Source Wavelet
  - theoretical formulation
  - current implementation and joint optimality conditions
- Robust and Sparsity Promoting Formulations
  - motivating application: basis pursuit denoise with **huber**
  - generalized analysis of the SPGℓ_1 framework
  - inexact Secant method

SPARSITY PROMOTION FOR GEOPHYSICS USING SPGℓ_1

Seismic Acquisition

Data are acquired in the field using thousands of source experiments. For every source experiment, time traces are obtained by an array of receivers.

Seismic surveys take months to complete, and full data volume is overwhelming.



Missing Traces Interpolation

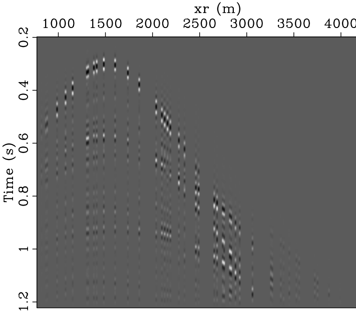
$$\text{BP}_\sigma: \quad \min \quad \|x\|_1 \quad \text{st} \quad \|b - RC^*x\|_2 \leq \boldsymbol{\sigma}$$

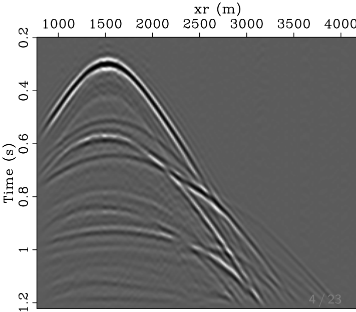
Problem Specification

b	seismic shot gathers with missing traces
C	Curvelet transform
R	Restriction to available data
x	curvelet coefficients of the data
σ	error level

Results

Missing traces can be recovered.





Sparse Formulation for Migration

$$\boxed{\mathrm{BP}_\sigma: \quad \min \quad \|x\|_1 \quad \text{st} \quad \|r - J^*C^*x\|_2 \leq \boldsymbol{\sigma}}$$

Problem Specification

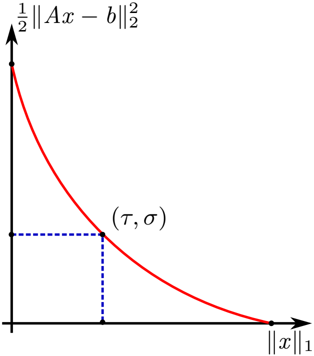
r	residual at smooth model estimate
m	smooth velocity estimate
J	Jacobian of forward model
C	Curvelet transform
x	curvelet coefficients of the update
σ	error level

Results

Improved recovery compared to LS inversion for subsampled data (Xiang Li's presentation).

Pareto curve: [Berg & F., 2008, 2011]

BP_σ:	$\min$	$\|x\|_1$	st	$\frac{1}{2}\|Ax-b\|_2^2 \le \boldsymbol{\sigma}$
LS_τ:	$\min$	$\frac{1}{2}\|Ax-b\|_2^2$	st	$\|x\|_1 \le \boldsymbol{\tau}$



The key is the value function

$$v(\tau) := \min_{\|x\|_1 \le \tau} \frac{1}{2}\|Ax-b\|_2^2$$

Algorithm

1. Evaluate $v(\tau)$ by solving LS_τ inexactly — projected gradient
2. Compute $v'(\tau)$ inexactly — duality theory
3. Solve $v(\tau) = \sigma$ — Inexact Newton's method

Root Finding: $v(\tau) = \sigma$

Approximately solve

minimize $\quad \frac{1}{2}\|Ax - b\|_2^2$

subj to $\quad \|x\|_1 \le \tau_k$

Approximate values

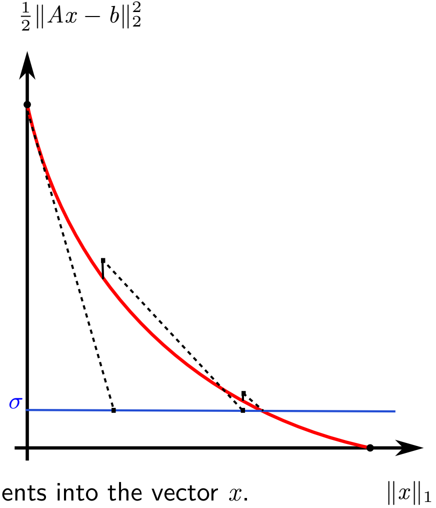
obtain approximate v_k, v'_k

Newton update

$\tau_{k+1} \leftarrow \tau_k - (v_k - \sigma)/v'_k$

Warm starts on iterates

Slowly include non-zero components into the vector x.



DIMENSIONALITY REDUCTION FOR BPDN

Structure of Seismic LASSO Problems

In least squares migration, the LASSO problem

$$\mathsf{LS}_\tau: \quad \min \quad \phi(x) := \tfrac{1}{2}\|Ax - b\|_2^2 \quad \text{st} \quad \|x\|_1 \le \boldsymbol{\tau}$$

has special structure:

$$\phi(x) = \frac{1}{2}\| \underbrace{\nabla\boldsymbol{\mathcal{F}}[\mathbf{m}_0;\mathbf{Q}]}_{A} \underbrace{\delta\mathbf{m}}_{x} - \underbrace{\delta\mathbf{d}}_{b} \|_F^2$$

Because of problem size, we want to work with dimensionality reduced problems:

$$\begin{aligned} \min \quad & \tilde{\phi}(x) := \tfrac{1}{2}\| \underbrace{\nabla\boldsymbol{\mathcal{F}}[\mathbf{m}_0;\widetilde{\mathbf{Q}}]}_{\widetilde{A}} \underbrace{\delta\mathbf{m}}_{x} - \underbrace{\widetilde{\delta\mathbf{D}}}_{\widetilde{b}} \|_F^2 \\ \text{st} \quad & \|x\|_1 \le \boldsymbol{\tau} \end{aligned}$$

where

$$\widetilde{\mathbf{Q}} = \mathbf{QW}\,, \quad \widetilde{\delta\mathbf{D}} = \delta\mathbf{DW}$$

for some random matrix $\mathbf{W}$. Subsampled LASSO problems use $\frac{k}{n}$ of the full volume, where k is the number of columns in $\mathbf{W}$.

Stochastic Root Finding: $v(\tau) = \sigma$

Approximately solve

$$\tilde{v}_k = \min \|\tilde{A}x - \tilde{b}\|_2^2$$
$$\text{subj to } \|x\|_1 \le \tau_k$$

Sampling Schemes

Easy to sample so that
$v_k = \|A\bar{x} - b\|_2^2 = \mathbb{E}[\|\tilde{A}\bar{x} - \tilde{b}\|_2^2$
and
$A^T(A\bar{x} - b) = \mathbb{E}[\tilde{A}^T(\tilde{A}\bar{x} - \tilde{b})]$

Approximate values

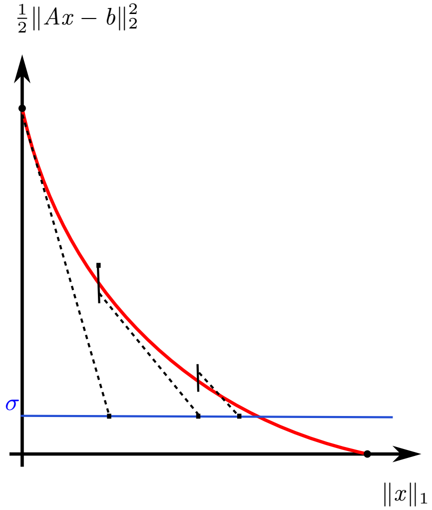
Therefore we have approx.
$v(\tau), v'(\tau)$.

Randomized Newton update

$$\tau_{k+1} \leftarrow \tau_k - (v_k - \sigma)/v'_k$$

Warm starts on iterates

Slowly include non-zero components into the vector x.



Conclusions and Questions

Conclusions:

- Dimensionality reduction techniques allow Least-Squares migration at a fraction of the cost of the full problem.
- The approach may be interpreted as a stochastic root-finding method, extending SPGℓ_1.

Open Questions:

- How close are random estimates of v_k to the inexact but nonrandom estimates of v_k?
- How can we control the noise so that the random error is dominated by the 'inexact' error?
- What can we say about the random estimates of $v'(\tau)$?
- What can we say about the solutions x_k obtained from the randomized LASSO problems?

SIMULTANEOUS ESTIMATION OF GREEN'S FUNCTION AND SOURCE WAVELET

A Different Way to Write EPSI:

Consider the EPSI formulation:

$$\min_{\mathbf{G},\mathbf{q}} \|\mathbf{G}\|_1 \quad \text{s.t.} \quad \|\mathbf{P} - \mathbf{G}\underbrace{(\mathrm{diag}(\mathbf{q}) + \mathbf{RP})}_{\mathbf{A}(\mathbf{q})}\|_F^2 \leq \sigma \tag{1}$$

where $\mathbf{G}$ is the Green's function, $\mathbf{q}$ is a vector representing the source wavelet, $\mathbf{P}$ is the recorded data, and $\mathbf{R}$ is the known reflection operator.

Note that (1) is an extension of the BPDN formulation. It is therefore natural to consider

$$\min_{\mathbf{G},\mathbf{q}} \phi(\mathbf{G},\mathbf{q}) := \|\mathbf{P} - \mathbf{G}\mathbf{A}(\mathbf{q})\|_F^2 \quad \text{s.t.} \quad \|\mathbf{G}\|_1 \leq \tau \tag{2}$$

Note that (2) is an extension to the standard LASSO problem.

Solving for optimal source wavelet

$$\min_{\mathbf{G},\mathbf{q}} \phi(\mathbf{G},\mathbf{q}) = \|\mathbf{P} - \mathbf{G}\mathbf{A}(\mathbf{q})\|_F^2 \quad \text{s.t.} \quad \|\mathbf{G}\|_1 \ \leq \tau$$

For fixed $\mathbf{G}$, it is easy to solve for an optimal $\mathbf{q}$:

$$q_i(\mathbf{G}) = \frac{(\mathbf{G}_{\mathbf{i},\cdot})^T(\mathbf{P} - \mathbf{GRP})_{\mathbf{i},\cdot}}{(\mathbf{G}_{\mathbf{i},\cdot})^T(\mathbf{G}_{\mathbf{i},\cdot})}$$

Note that this looks a lot like the formula for LS source estimation.

The upshot is that if at every iteration we solve for the optimal $\mathbf{q}$ as a function of $\mathbf{G}$, we can say something about convergence.

Spectral Projected Variable Projection?

Suppose we think of $\hat{\mathbf{q}} = \mathbf{q}(\mathbf{G})$ as a function of $\mathbf{G}$, either available in closed form or quickly obtainable. Then the modified LASSO problem can be rewritten

$$\min_{\mathbf{G}} f(\mathbf{G}) = \phi(\mathbf{G}, \hat{\mathbf{q}}) = \|\mathbf{P} - \mathbf{G}\mathbf{A}(\hat{\mathbf{q}})\|_F^2 \quad \text{s.t.} \quad \|\mathbf{G}\|_1 \le \tau$$

The key point is that the gradient ∇f is given by

$$2\mathbf{A}(\hat{\mathbf{q}})(\mathbf{P} - \mathbf{G}\mathbf{A}(\hat{\mathbf{q}})) + \mathbf{0}$$

Therefore, the Spectral Gradient algorithm will converge to a local minimum of $f(\mathbf{G})$, as long as $\hat{\mathbf{q}}$ is updated before every gradient computation.

Moreover, when it does, we have an optimal $\bar{\mathbf{G}}$ and the corresponding $\mathbf{q}(\bar{\mathbf{G}})$, so we have solved $\phi(\mathbf{G}, \mathbf{q})$!

Conclusions and Questions

Conclusions:

- EPSI is a true extension of the BPDN framework, and can be attacked using the same tools.
- At least for the extended LASSO problems, it is clear that we can find a local minimum.

Open Questions:

- Currently, the EPSI implementation does not update $\mathbf{q}$ before every gradient computation. Will this change make a difference?
- What does the 'Pareto' curve for $f(\mathbf{G})$ look like? Is it convex?
- Can we prove that we solve the extended BPDN problem?

GENERALIZED FRAMEWORK FOR ROBUST AND SPARSE FORMULATIONS

Robust Formulations

Statistical interpretation of least-squares problems:

Objective	Statistical Model	**Error model**
$\min_x \|Ax - b\|_2^2$	$b = Ax + \epsilon$	$\epsilon \sim N(0, I)$.

Formulations based on least squares recovery and Gaussian modeling are known to be **vulnerable to outliers** in the data, which may be random errors or **systematic features unexplained by the forward model**.

Robust formulations can be derived by replacing the Gaussian distribution assumption on ϵ with a heavier-tailed distribution:

Formulation	**Objective**	**Error model**
Gaussian	$\sum (a_i^T x - b_i)^2$	$\epsilon_i \sim N(0, 1)$
Laplace	$\sum \|a_i^T x - b_i\|$	$\epsilon_i \sim L(0, 1)$
Huber	$\sum \rho_h(a_i^T x - b_i)$	$\epsilon_i \sim H(0, 1)$
Student's t	$\sum \log(k + (a_i^T x - b_i)^2)$	$\epsilon_i \sim S(0, 1)$

Robust References

Robust methods are an active area of research in statistics and across many applications. A few references:

Robust Statistics

Huber '81	M-estimates
Marona et al. '06	M-estimates and other topics
Seber and Wild, '00	Robust nonlinear regression

Seismic Imaging

Guitton and Symes, '03	Huber norm
Bube and Nemeth, '07	Huber and hybrid ℓ_1-ℓ_2 norms
Brossier et al. '09 and '10	Huber and ℓ_1 norms
AYA, MPF, FJH, TVL. '11	Dim. red. & robust penalties (Student's t).

We focus on the convex robust formulations in this talk. Can we combine these ideas with sparsity promotion via BPDN?

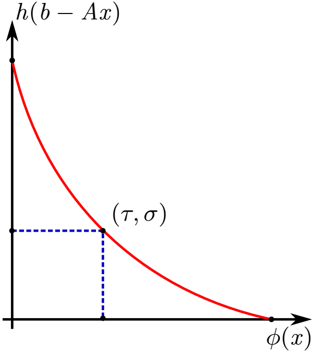
Generalized Objectives, Penalties, and Value Functions

$\mathcal{P}_1(\sigma)$:	$\min$	$\phi(x)$	st	$h(b - Ax) \le \boldsymbol{\sigma}$
$\mathcal{P}_2(\tau)$:	$\min$	$h(b - Ax)$	st	$\phi(x) \le \boldsymbol{\tau}$

Problems $\mathcal{P}_1(\sigma)$ and $\mathcal{P}_2(\tau)$ are linked by

$$v_2(\tau) := \min_{\phi(x) \le \tau} h(b - Ax)$$

$v_2(\tau)$ describes tradeoff between misfit measure h and regularization function ϕ. *Note that h can be robust!*



Broad summary of results:

1. $v_2(\tau)$ is always convex, but may not be differentiable.
2. As of last week, we completely understand variational properties of $v_2(\tau)$ with respect to b and τ for **arbitrary convex** ϕ, h.
3. Solving $v_2(\tau) = \sigma$ can always be done using an inexact secant method.

A Little Less General

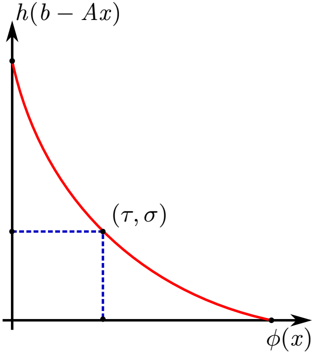
$$\boxed{\begin{array}{llll} \mathcal{P}_1(\sigma): & \min & \|x\|_1 & \text{st} \quad h(b-Ax) \le \boldsymbol{\sigma} \\ \mathcal{P}_2(\tau): & \min & h(b-Ax) & \text{st} \quad \|x\|_1 \le \boldsymbol{\tau} \end{array}}$$

$$v_2(\tau) := \min_{\|x\|_1 \le \tau} h(b - Ax)$$

Suppose that h is differentiable. Then

$$v_2'(\tau) = -\|A^T \nabla h(b - A\bar{x})\|_\infty ,$$

where $\bar{x}$ is the minimizing parameter value.



Some examples of convex differentiable h:

1. Huber
2. Hybrid (Bube and Nemeth; strictly convex robust objective)

In some cases, $v_2(\tau)$ is not differentiable, but it is always convex. We therefore introduce a novel inexact secant method.

Inexact Secant method for $v_2(\tau) = \sigma$

Theorem

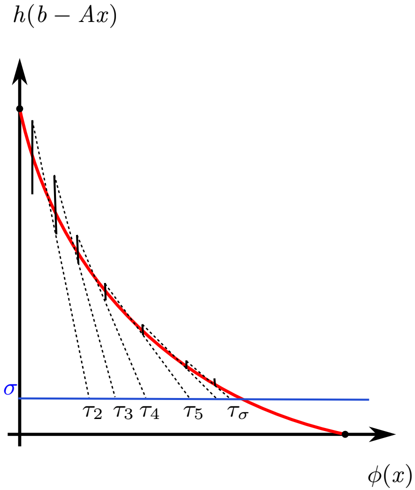
The inexact secant method for finding $v_2(\tau) = \sigma$, given by

$$\tau_{k+1} \leftarrow \tau_k - \frac{l(\tau_k) - \sigma}{m_k}$$

$$m_k = \frac{l(\tau_k) - u(\tau_{k-1})}{(\tau_k - \tau_{k-1})}$$

is superlinearly convergent as long as

1. $u(\tau_k) - l(\tau_k)$ shrinks fast enough
2. $v_2(\tau)$ has a nonzero subgradient at τ_σ.



Sparse and Robust Formulation

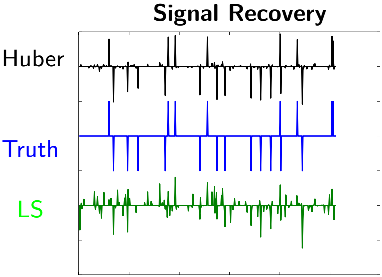
$$\boxed{\text{HBP}_\sigma: \quad \min \quad \|x\|_1 \quad \text{st} \quad \rho(d - Ax) \leq \boldsymbol{\sigma}}$$

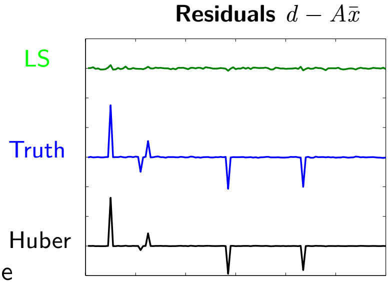
Problem Specification

x	20-sparse spike train in $\mathbb{R}^{512}$
d	measurements in $\mathbb{R}^{120}$
A	Measurement matrix satisfying RIP
ρ	Huber function
σ	error level set at $.01$

Results

In the presence of outliers, the robust formulation recovers the spike train, while the standard formulation does not.





Conclusions and Questions

Conclusions:

- We have shown an extension of the foundation underlying SPGℓ_1 that allows many potential new applications.
- The motivating application, combining robustness and sparsity, requires only minor modifications to an existing SPGℓ_1 implementation. In fact, Tim has already implemented it in SPGℓ_1.

Open Questions:

- How much improvement can we get for real applications using the Huber-ℓ_1 formulation?
- What other applications within the general framework presented are interesting and feasible?
- What do we do when either the misfit function h or the regularizer ϕ is non-convex?

SLIM

Acknowledgements

This work was in part financially supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (22R81254) and the Collaborative Research and Development Grant DNOISE II (375142-08). This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BP, Chevron, BGP, ConocoPhillips, PGS, Petrobras, Total SA, and WesternGeco.