Iteratively Reweighted ℓ_1 Approaches to Sparse Composite Regularization

Phil Schniter

THE OHIO STATE UNIVERSITY

Joint work with **Prof. Rizwan Ahmad** (OSU)

Supported in part by NSF grant CCF-1018368.

MATHEON Conf. on Compressed Sensing and its Applications
TU-Berlin — Dec 11, 2015

Outline

Introduction

- Goal: Recover signal $\boldsymbol{x} \in \mathbb{C}^N$ from noisy linear measurements

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{w} \in \mathbb{C}^M$$

where usually $M \ll N$.

- Approach: Solve the optimization problem

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \gamma \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + R(\boldsymbol{x}),$$

with $\gamma > 0$ controlling the measurement fidelity.

- Question: How should we choose penalty/regularization $R(\boldsymbol{x})$?

Typical Choices of Penalty

Say $\boldsymbol{\Psi x}$ is (approximately) sparse for "analysis operator" $\boldsymbol{\Psi} \in \mathbb{C}^{L \times N}$.

ℓ_0 penalty: $R(\boldsymbol{x}) = \|\boldsymbol{\Psi x}\|_0$

- Impractical: optimization problem is NP hard

ℓ_1 penalty (generalized LASSO): $R(\boldsymbol{x}) = \|\boldsymbol{\Psi x}\|_1$

- Tightest convex relaxation of ℓ_0 penalty
- Fast algorithms: ADMM, MFISTA, NESTA-UP, grAMPa ...

non-convex penalties

- $R(\boldsymbol{x}) = \|\boldsymbol{\Psi x}\|_p$ for $p \in (0,1)$ (via IRW-L2)
- $R(\boldsymbol{x}) = \sum_{l=1}^{L} \log(\epsilon + |\boldsymbol{\psi}_l^{\mathsf{T}} \boldsymbol{x}|)$ with $\epsilon \geq 0$ (via IRW-L1)
- many others...

Choice of Analysis Operator

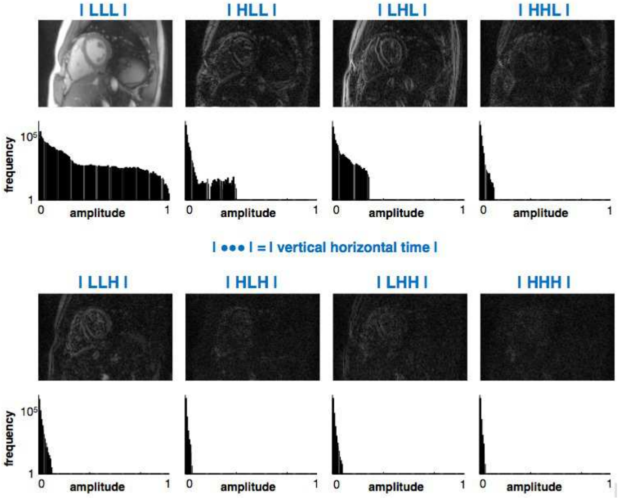
How to choose $\boldsymbol{\Psi}$ in practice?

- Maybe a wavelet transform? Which one?
- Maybe a concatenation of several transforms $\begin{bmatrix} \boldsymbol{\Psi}_1 \\ \vdots \\ \boldsymbol{\Psi}_D \end{bmatrix}$ (e.g., SARA[1])?
- What if signal is more sparse in one dictionary than another?
  Can we compensate for this?
  Can we exploit this?

[1] Carrillo, McEwen, Van De Ville, Thiran, Wiaux, "Sparsity averaged reweighted analysis," *IEEE SPL*, 2013

Example: Undecimated Wavelet Transform of MRI Cine

Note different sparsity rate in each subband of 1-level UWT:



Composite ℓ_1 Penalties

- We propose to use composite ℓ_1 (Co-L1) penalties of the form

$$R(\boldsymbol{x};\boldsymbol{\lambda}) \triangleq \sum_{d=1}^{D} \lambda_d \|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1, \quad \lambda_d \geq 0$$

  where $\boldsymbol{\Psi}_d \in \mathbb{C}^{L_d \times N}$ have unit-norm rows.

- The $\boldsymbol{\Psi}_d$ could be chosen, for example, as
  - different DWTs (i.e., db1,db2,db3,. . . ,db10),
  - different subbands of a given DWT,
  - row-subsets of $\boldsymbol{I}$ (i.e., group/hierarchical sparsity), or
  - all of the above.

- We then aim to simultaneously tune the weights $\{\lambda_d\}$ and recover the signal $\boldsymbol{x}$.

The Co-L1 Algorithm

input: $\{\boldsymbol{\Psi}_d\}_{d=1}^{D}$, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\gamma > 0$, $\epsilon \geq 0$
if $\boldsymbol{\Psi}_d \boldsymbol{x} \in \mathbb{R}^{L_d}$ then $C_d = 1$, elseif $\boldsymbol{\Psi}_d \boldsymbol{x} \in \mathbb{C}^{L_d}$ then $C_d = 2$.
initialization: $\lambda_d^{(1)} = 1 \;\forall d$
for $t = 1, 2, 3, \ldots$

$$\boldsymbol{x}^{(t)} \leftarrow \arg\min_{\boldsymbol{x}} \left\{ \gamma \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \sum_{d=1}^{D} \lambda_d^{(t)} \|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1 \right\}$$

$$\lambda_d^{(t+1)} \leftarrow \frac{C_d L_d}{\epsilon + \|\boldsymbol{\Psi}_d \boldsymbol{x}^{(t)}\|_1}, \quad d = 1, \ldots, D$$

end
output: $\boldsymbol{x}^{(t)}$

- leverages existing ℓ_1 solvers (e.g., ADMM, MFISTA, NESTA-UP, grAMPa),
- reduces to the IRW-L1 algorithm [Figueiredo,Nowak'07] when $L_d = 1 \;\forall d$ (single-atom dictionaries).
- applies to both real- and complex-valued cases,

The Co-IRW-L1 Algorithm

input: $\{\boldsymbol{\Psi}_d\}_{d=1}^{D}$, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\gamma > 0$

initialization: $\lambda_d^{(1)} = 1 \ \forall d$, $\boldsymbol{W}_d^{(1)} = \boldsymbol{I} \ \forall d$

for $t = 1, 2, 3, \dots$

$$\boldsymbol{x}^{(t)} \leftarrow \arg\min_{\boldsymbol{x}} \left\{ \gamma \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \sum_{d=1}^{D} \lambda_d^{(t)} \|\boldsymbol{W}_d^{(t)} \boldsymbol{\Psi}_d \boldsymbol{x}\|_1 \right\}$$

$$(\lambda_d^{(t+1)}, \epsilon_d^{(t+1)}) \leftarrow \arg \max_{\lambda_d \in \Lambda, \epsilon_d > 0} \log p(\boldsymbol{x}^{(t)}; \boldsymbol{\lambda}, \boldsymbol{\epsilon}), \ d = 1, \dots, D$$

$$\boldsymbol{W}_d^{(t+1)} \leftarrow \operatorname{diag}\left\{ \frac{1}{\epsilon_d^{(t+1)} + |\boldsymbol{\psi}_{d,1}^{\mathsf{T}} \boldsymbol{x}^{(t)}|}, \cdots, \frac{1}{\epsilon_d^{(t+1)} + |\boldsymbol{\psi}_{d,L_d}^{\mathsf{T}} \boldsymbol{x}^{(t)}|} \right\}, d = 1, \dots, D$$

end

output: $\boldsymbol{x}^{(t)}$

- tunes both λ_d and diagonal $\boldsymbol{W}_d$ for all d: hierarchical weighting.
- also tunes regularization parameters ϵ_d for all d.

Understanding Co-L1 and Co-IRW-L1

In the sequel, we provide four interpretations of each algorithm:

1. Majorization-minimization (MM) for a particular non-convex penalty,
2. a particular approximation of ℓ_0 minimization,
3. Bayesian estimation according to a particular hierarchical prior,
4. variational EM algorithm under a particular prior.

Outline

Optimization Interpretations of Co-L1

Co-L1 is an MM approach to the weighted log-sum optimization problem

$$\arg\min_{\boldsymbol{x}} \left\{ \gamma \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \sum_{d=1}^{D} L_d \log(\epsilon + \|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1) \right\}$$

and

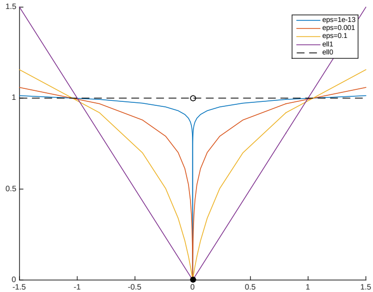
As $\epsilon \to 0$, Co-L1 aims to solve the weighted $\ell_{1,0}$ problem

$$\arg\min_{\boldsymbol{x}} \left\{ \gamma \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \sum_{d=1}^{D} L_d \, 1_{\|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1 > 0} \right\}$$

Note: L_d is # atoms in dictionary $\boldsymbol{\Psi}_d$, and $1_{\square}$ is the indicator function.

Approximate-ℓ_0 Interpretation of Log-Sum Penalty

$$\frac{1}{\log(1/\epsilon)}\sum_{n=1}^{N}\log(\epsilon+|u_n|)$$
$$=\frac{1}{\log(1/\epsilon)}\Bigg[\sum_{n:\,x_n=0}\log(\epsilon)$$
$$+\sum_{n:\,x_n\neq 0}\log(\epsilon+|u_n|)\Bigg]$$
$$=\|\boldsymbol{x}\|_0 - N + \frac{\sum_{n:\,x_n\neq 0}\log(\epsilon+|u_n|)}{\log(1/\epsilon)}$$



As $\epsilon \to 0$, the log-sum penalty becomes a scaled and shifted version of the ℓ_0 penalty.

Bayesian Interpretations of Co-L1

Co-L1 is an MM approach to Bayesian MAP estimation under an AWGN likelihood and the hierarchical prior

$$p(\boldsymbol{x}|\boldsymbol{\lambda}) = \prod_{d=1}^{D} \left(\frac{\lambda_d}{2}\right)^{L_d} \exp\left(-\lambda_d \|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1\right) \quad \text{i.i.d. Laplacian}$$

$$p(\boldsymbol{\lambda}) = \prod_{d=1}^{D} \Gamma\left(0, \frac{1}{\epsilon}\right), \quad \text{i.i.d. Gamma (i.i.d. Jeffrey's as } \epsilon \to 0)$$

and

As $\epsilon \to 0$, Co-L1 is a variational EM approach to estimating (deterministic) $\boldsymbol{\lambda}$ under an AWGN likelihood and the prior

$$p(\boldsymbol{x}; \boldsymbol{\lambda}) = \prod_{d=1}^{D} \left(\frac{\lambda_d}{2}\right)^{L_d} \exp\left(-\lambda_d(\|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1 + \epsilon)\right) \quad \text{i.i.d. Laplacian as } \epsilon \to 0$$

Outline

A Simplified Version of Co-IRW-L1

Consider the real-valued and fixed-ϵ_d variant of Co-IRW-L1.

1: input: $\{\boldsymbol{\Psi}_d\}_{d=1}^D$, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\gamma > 0$, $\epsilon_d > 0\ \forall d$

2: initialization: $\lambda_d^{(1)} = 1\ \forall d$, $\boldsymbol{W}_d^{(1)} = \boldsymbol{I}\ \forall d$

3: for $t = 1, 2, 3, \dots$

4: $$\boldsymbol{x}^{(t)} \leftarrow \arg\min_{\boldsymbol{x}} \left\{ \gamma \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \sum_{d=1}^{D} \lambda_d^{(t)} \|\boldsymbol{W}_d^{(t)} \boldsymbol{\Psi}_d \boldsymbol{x}\|_1 \right\}$$

5: $$\lambda_d^{(t+1)} \leftarrow \left[\frac{1}{L_d} \sum_{l=1}^{L_d} \log \left(1 + \frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}} \boldsymbol{x}^{(t)}|}{\epsilon_d} \right) \right]^{-1} + 1, \quad d = 1, \dots, D$$

6: $$\boldsymbol{W}_d^{(t+1)} \leftarrow \operatorname{diag} \left\{ \frac{1}{\epsilon_d + |\boldsymbol{\psi}_{d,1}^{\mathsf{T}} \boldsymbol{x}^{(t)}|}, \cdots, \frac{1}{\epsilon_d + |\boldsymbol{\psi}_{d,L_d}^{\mathsf{T}} \boldsymbol{x}^{(t)}|} \right\}, \ d = 1, \dots, D,$$

7: end

8: output: $\boldsymbol{x}^{(t)}$

Optimization Interpretations of real-Co-IRW-L1-ϵ

Real-Co-IRW-L1-ϵ is an MM approach to the non-convex optimization

$$\arg\min_{\boldsymbol{x}} \left\{ \gamma \|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2^2 + \sum_{d=1}^{D} \sum_{l=1}^{L_d} \log \left[\left(\epsilon_d + |\boldsymbol{\psi}_{d,l}^{\mathsf{T}} \boldsymbol{x}| \right) \sum_{i=1}^{L_d} \log \left(1 + \frac{|\boldsymbol{\psi}_{d,i}^{\mathsf{T}} \boldsymbol{x}|}{\epsilon_d} \right) \right] \right\}$$

and

As $\epsilon_d \to 0$, real-Co-IRW-L1-ϵ aims to solve the ℓ_0+ weighted $\ell_{0,0}$ problem

$$\arg\min_{\boldsymbol{x}} \left\{ \gamma \|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2^2 + \|\boldsymbol{\Psi x}\|_0 + \sum_{d=1}^{D} L_d \, 1_{\|\boldsymbol{\Psi}_d \boldsymbol{x}\|_0 > 0} \right\}$$

Note: L_d is the size of dictionary $\boldsymbol{\Psi}_d$, and $1_{\square}$ is the indicator function.

Bayesian Interpretations of real-Co-IRW-L1-ϵ

Real-Co-IRW-L1 is an MM approach to Bayesian MAP estimation under an AWGN likelihood and the hierarchical prior

$$p(\boldsymbol{x}|\boldsymbol{\lambda}) = \prod_{d=1}^{D}\prod_{l=1}^{L_d} \frac{\lambda_d}{2\epsilon_d}\left(1 + \frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|}{\epsilon_d}\right)^{-(\lambda_d+1)} \quad \text{i.i.d. generalized-Pareto}$$

$$p(\boldsymbol{\lambda}) = \prod_{d=1}^{D} p(\lambda_d), \quad p(\lambda_d) \propto \begin{cases} \frac{1}{\lambda_d} & \lambda_d > 0 \\ 0 & \text{else} \end{cases} \quad \text{Jeffrey's non-informative}$$

and

Real-Co-IRW-L1 is a variational EM approach to estimating (deterministic) $\boldsymbol{\lambda}$ under an AWGN likelihood and the prior

$$p(\boldsymbol{x};\boldsymbol{\lambda}) = \prod_{d=1}^{D}\prod_{l=1}^{L_d} \frac{\lambda_d - 1}{2\epsilon_d}\left(1 + \frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|}{\epsilon_d}\right)^{-\lambda_d} \quad \text{i.i.d. generalized-Pareto}$$

The Co-IRW-L1 Algorithm

Finally, we self-tune $\epsilon_d \; \forall d$ and allow for real or complex quantities:

1: input: $\{\boldsymbol{\Psi}_d\}_{d=1}^D$, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\gamma > 0$

2: if $\boldsymbol{\Psi x} \in \mathbb{R}^L$, use $\Lambda = (1, \infty)$ and the real version of $\log p(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\epsilon})$;
elseif $\boldsymbol{\Psi x} \in \mathbb{C}^L$, use $\Lambda = (2, \infty)$ and the complex version of $\log p(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\epsilon})$.

3: initialization: $\lambda_d^{(1)} = 1 \; \forall d$, $\boldsymbol{W}_d^{(1)} = \boldsymbol{I} \; \forall d$

4: for $t = 1, 2, 3, \ldots$

5: $$\boldsymbol{x}^{(t)} \leftarrow \arg\min_{\boldsymbol{x}} \left\{ \gamma \|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2^2 + \sum_{d=1}^{D} \lambda_d^{(t)} \|\boldsymbol{W}_d^{(t)} \boldsymbol{\Psi}_d \boldsymbol{x}\|_1 \right\}$$

6: $$(\lambda_d^{(t+1)}, \epsilon_d^{(t+1)}) \leftarrow \arg \max_{\lambda_d \in \Lambda, \epsilon_d > 0} \log p(\boldsymbol{x}^{(t)}; \boldsymbol{\lambda}, \boldsymbol{\epsilon}), \; d = 1, ..., D$$

7: $$\boldsymbol{W}_d^{(t+1)} \leftarrow \operatorname{diag} \left\{ \frac{1}{\epsilon_d^{(t+1)} + |\boldsymbol{\psi}_{d,1}^\mathsf{T} \boldsymbol{x}^{(t)}|}, \cdots, \frac{1}{\epsilon_d^{(t+1)} + |\boldsymbol{\psi}_{d,L_d}^\mathsf{T} \boldsymbol{x}^{(t)}|} \right\}, \; d = 1, ..., D$$

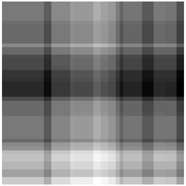
8: end

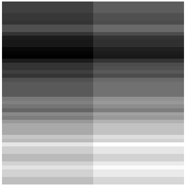
9: output: $\boldsymbol{x}^{(t)}$

Outline

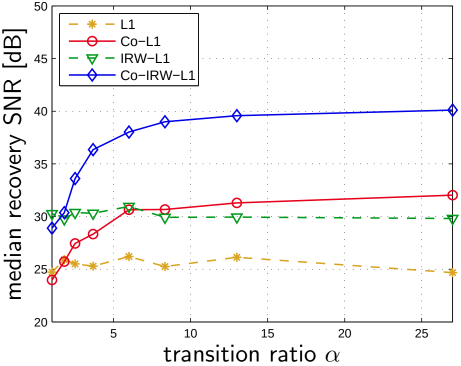
Experiment: Synthetic finite difference image

$\alpha = 1$ $\alpha = 27$





- 48×48 image with a total of 28 horiz & vert transitions.
- $\alpha \triangleq \frac{\#\text{ vertical transitions}}{\#\text{ horizontal transitions}}$
- $\mathbf{\Psi}_1 =$ vertical finite difference, $\mathbf{\Psi}_2 =$ horizon. finite difference
- "spread-spectrum" $\mathbf{\Phi}$
- sampling ratio $\frac{M}{N} = 0.3$
- AWGN @ 30 dB SNR

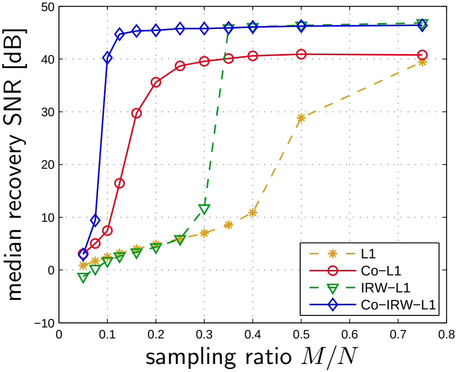


$\Rightarrow$ The composite algorithms significantly outperform the non-composite ones

$\Rightarrow$ Performance improves as sparsities become more disparate!

Experiment: Shepp-Logan Phantom

- 96×96 image
- $\mathbf{\Psi} \in \mathbb{R}^{7N \times N}$ = 2D UWT-db1, $\mathbf{\Psi}_d \in \mathbb{R}^{N \times N}\ \forall d$
- "spread-spectrum" $\mathbf{\Phi}$
- AWGN @ 30 dB SNR

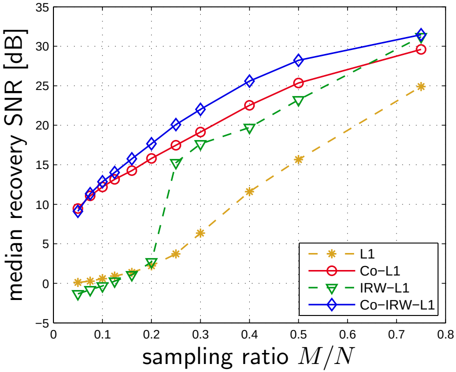


⇒ The composite algorithms significantly outperform the non-composite ones

⇒ Performance gap is larger for small M/N

Experiment: Cameraman

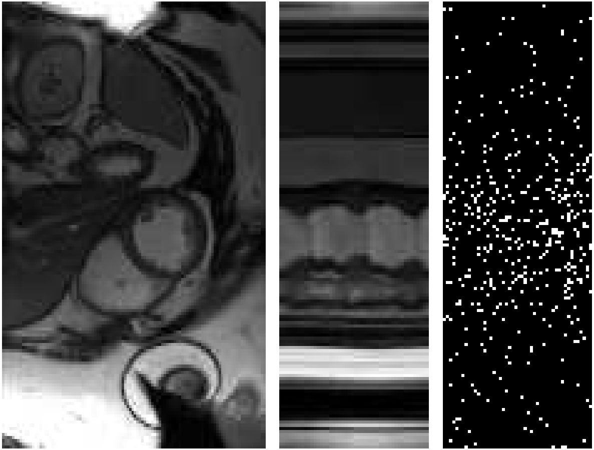
- 96×104 image
- $\mathbf{\Psi} \in \mathbb{R}^{7N \times N}$ = 2D UWT-db1, $\mathbf{\Psi}_d \in \mathbb{R}^{N \times N}\ \forall d$
- "spread-spectrum" $\mathbf{\Phi}$
- AWGN @ 40 dB SNR



$\Rightarrow$ The composite algorithms significantly outperform the non-composite ones

$\Rightarrow$ Performance gap is larger for small M/N

Experiment: 1D Dynamic MRI

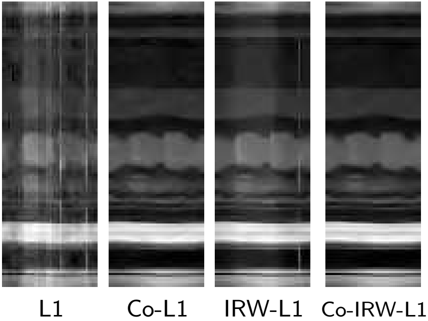


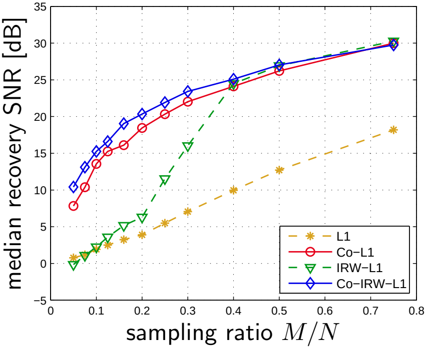
x-y profile x-t profile k-t sampling

- 144×48 spatiotemporal profile extracted from MRI cine
- $\mathbf{\Psi} \in \mathbb{R}^{3N \times N}$: [db1;db2;db3] 2D DWT
- $\mathbf{\Phi}$: variable density random Fourier
- AWGN @ 30 dB SNR

Experiment: 1D Dynamic MRI (cont.)

sampling ratio $M/N = 0.3$





- The composite algs significantly outperform the non-composite ones at small measurement ratios M/N
- Little advantage to Co-IRW-L1 over Co-L1 in this experiment

Average Runtimes for Previous Experiments

	Shepp-Logan	Cameraman	dMRI
L1	8.12s	9.88s	22.0s
Co-L1	8.83s	12.8s	21.7s
IRW-L1	7.95s	12.7s	24.1s
Co-IRW-L1	9.29s	16.9s	29.6s

The composite algs run only $1.3\times$ longer than the non-composite ones.

Open Questions

- Performance guarantees?
- Convergence guarantees? (So far we have only established an asymptotic stationary point condition using an MM analysis of Julien Mairal.[2]
- Design of dictionaries $\{\boldsymbol{\Psi}_d\}$?
- Extension to matrix compressive sensing (e.g., low-rank, row-sparse, column-sparse, etc.)?

[2] J. Mairal, "Optimization with first-order surrogate functions," *ICML*, 2013.

Conclusions

- We proposed a new "composite-L1" approach to L2-penalized signal reconstruction that learns and exploits differences in sparsity across sub-dictionaries.
- Relative to standard L1 methods, our composite L1 methods give significant improvements in reconstruction SNR at low sampling rates, at the cost of very mild complexity increase.
- Our algorithms can be interpreted as MM approaches to non-convex optimization, approximate ℓ_0 methods, Bayesian methods, and variational Bayesian methods.

References

Thanks!

1. R. Ahmad and P. Schniter, "Iteratively Reweighted L1 Approaches to Sparse Composite Regularization," *IEEE Transactions on Computational Imaging*, to appear. (See also http://arxiv.org/abs/1504.05110v4)
2. R. Ahmad and P. Schniter, "Iteratively Reweighted L1 Approaches to L2-Constrained Sparse Composite Regularization," (See http://arxiv.org/abs/1504.05110v2)