# Regularizing inverse problems using sparsity-based signal models

Jeffrey A. Fessler
William L. Root Professor of EECS

EECS Dept., BME Dept., Dept. of Radiology
University of Michigan

http://web.eecs.umich.edu/~fessler

Work with Sai Ravishankar and Raj Nadakuditi

CSP Seminar

31 Mar. 2016

# Disclosure

- Research support from GE Healthcare
- Supported in part by NIH grants P01 CA-87634, U01 EB018753
- Equipment support from Intel Corporation

# Outline

**Why**

- Low-dose X-ray CT imaging
- Accelerated MR imaging
- Other inverse problems

**How**

- MAP estimation for inverse problems
- Classical regularization methods (some sparsity based)
- Contemporary regularization methods (all sparsity based)

# Outline

Why

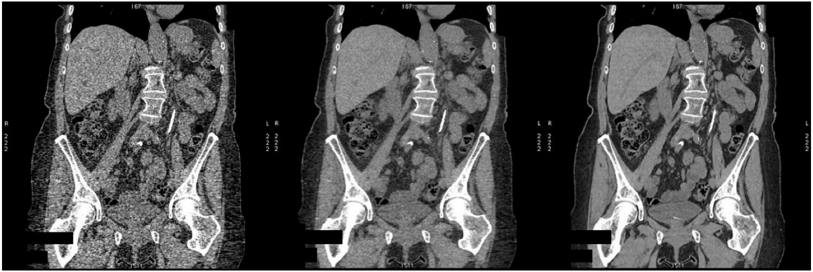
- Low-dose X-ray CT imaging
- Accelerated MR imaging
- Other inverse problems

How

- MAP estimation for inverse problems
- Classical regularization methods (some sparsity based)
- Contemporary regularization methods (all sparsity based)

# Improving X-ray CT image reconstruction

- A picture is worth 1000 words
- (and perhaps several 1000 seconds of computation?)



| Thin-slice FBP | ASIR (denoise) | Statistical |
|---|---|---|
| Seconds | A bit longer | Much longer |

Today's talk: less about computation, more about image quality

Right image used edge-preserving regularization

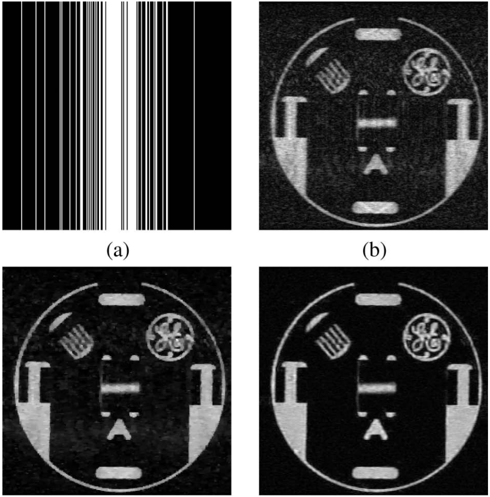
# Accelerating MR imaging

(a) 4× under-sampled MR k-space

(b) zero-filled reconstruction

(c) "compressed sensing" reconstruction with TV regularization

(d) adaptive dictionary learning regularization [1, Fig. 10]



(a) (b) (c) (d)

# Other ill-posed inverse problems

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \varepsilon$$

- compressed sensing ($\boldsymbol{A}$ random, wide)
- deblurring (restoration) ($\boldsymbol{A}$ Toeplitz, wide?)
- in-painting ($\boldsymbol{A}$ subset of rows of $\boldsymbol{I}$)
- denoising (not ill posed) ($\boldsymbol{A} = \boldsymbol{I}$)

# Outline

Why

- Low-dose X-ray CT imaging
- Accelerated MR imaging
- Other inverse problems

How

- MAP estimation for inverse problems
- Classical regularization methods (some sparsity based)
- Contemporary regularization methods (all sparsity based)

# Outline

Why

- Low-dose X-ray CT imaging
- Accelerated MR imaging
- Other inverse problems

How

- MAP estimation for inverse problems
- Classical regularization methods (some sparsity based)
- Contemporary regularization methods (all sparsity based)

# Inverse problems via MAP estimation

Unknown image $\boldsymbol{x}$ $\rightarrow$ [System model $\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x})$] $\rightarrow$ Data $\boldsymbol{y}$ $\rightarrow$ [Estimator] $\rightarrow$ $\hat{\boldsymbol{x}}$

If we have a prior $\mathrm{p}(\boldsymbol{x})$, then the MAP estimate is:

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y}) = \arg\max_{\boldsymbol{x}} \log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}) + \log \mathrm{p}(\boldsymbol{x}) .$$

For gaussian measurement errors and linear model:

$$-\log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}) \equiv \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{W}}^2$$

where $\|\boldsymbol{y}\|_{\boldsymbol{W}}^2 = \boldsymbol{y}' \boldsymbol{W} \boldsymbol{y}$ and $\boldsymbol{W}^{-1} = \mathrm{Cov}\{\boldsymbol{y} \mid \boldsymbol{x}\}$ is known

($\boldsymbol{A}$ from physics, $\boldsymbol{W}$ from statistics)

# Priors for MAP estimation

- If all images $\boldsymbol{x}$ are "plausible" (have non-zero probability) then

$$\mathrm{p}(\boldsymbol{x}) \propto \mathrm{e}^{-\mathrm{R}(\boldsymbol{x})} \implies -\log \mathrm{p}(\boldsymbol{x}) \equiv \mathrm{R}(\boldsymbol{x})$$

  (from fantasy / imagination / wishful thinking)

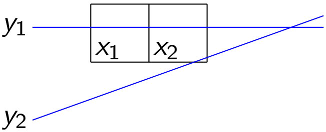
- MAP $\equiv$ regularized weighted least-squares (WLS) estimation:

$$\begin{aligned} \hat{\boldsymbol{x}} &= \arg\max_{\boldsymbol{x}} \; \log \mathrm{p}(\boldsymbol{y} \,|\, \boldsymbol{x}) + \log \mathrm{p}(\boldsymbol{x}) \\ &= \arg\min_{\boldsymbol{x}} \; \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{W}}^2 + \mathrm{R}(\boldsymbol{x}) \end{aligned}$$

- A regularizer $\mathrm{R}(\boldsymbol{x})$, aka log prior, is essential for high-quality solutions to ill-conditioned / ill-posed inverse problems.
- Why ill-posed? High ambitions...

# Example of ill-conditioned inverse problem

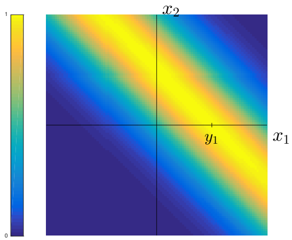
Two-pixel, two-ray "X-ray tomography" model:




$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 0 & 0.1 \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

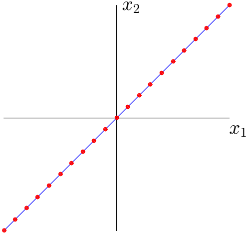
$\text{cond}(\boldsymbol{A}'\boldsymbol{A}) \approx 400$ $\qquad$ $\boldsymbol{A}$ is (roughly) square - somewhat typical

log-likelihood $\log \text{p}(\boldsymbol{y}|\boldsymbol{x})$:



# Subspace model: Alternative to regularization

Assuming $\boldsymbol{x}$ lies in a sufficiently low-dimensional subspace could make an inverse problem well conditioned.



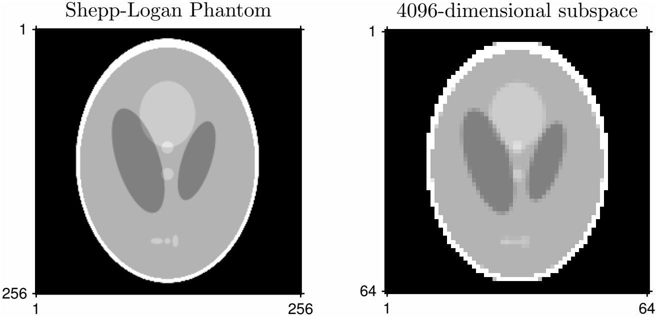

Assume $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{z}$ where $\boldsymbol{D} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\boldsymbol{z} \in \mathbb{R}^1$

($\boldsymbol{z}$ has only one nonzero element so very sparse!?)
Estimate coefficient(s): $\hat{\boldsymbol{z}} = \arg\min_{\boldsymbol{z}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{D}\boldsymbol{z}\|_2^2$, then $\hat{\boldsymbol{x}} = \boldsymbol{D}\hat{\boldsymbol{z}}$,

where $\boldsymbol{B} \triangleq \boldsymbol{A}\boldsymbol{D} = \begin{bmatrix} 2 \\ 0.1 \end{bmatrix}$ and $\operatorname{cond}(\boldsymbol{B}'\boldsymbol{B}) = 1$ which is perfect!

# Why not use subspace models?

Candès and Romberg (2005) [2] used 22 (noiseless) projection views, each with 256 samples.
$22 \cdot 256 = 5632$ measured values, vs $256^2 = 65536$ unknown pixels




Subspace representation (using pixel basis) is undesirably coarse.

# Outline

Why

- Low-dose X-ray CT imaging
- Accelerated MR imaging
- Other inverse problems

How

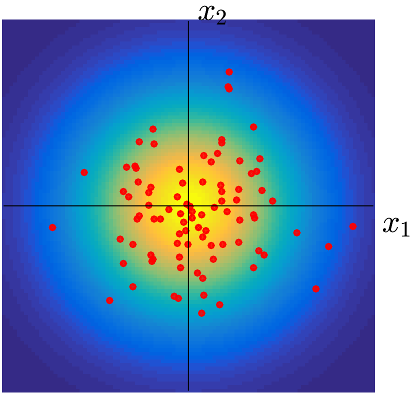
- MAP estimation for inverse problems
- Classical regularization methods (some sparsity based)
- Contemporary regularization methods (all sparsity based)

# Classical regularizers

- Tikhonov regularization (IID gaussian prior)
- Roughness penalty (Basic MRF prior)
- Sparsity in ambient space
- Edge-preserving regularization
- Total-variation (TV) regularization
- Black-box denoiser like NLM

# Tikhonov regularization

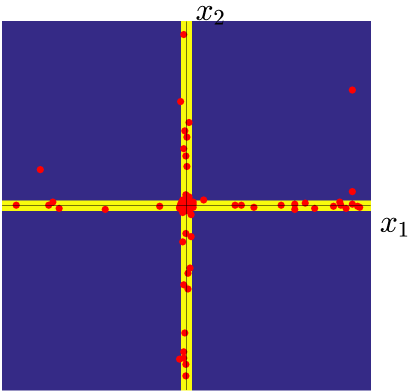
$$R(\boldsymbol{x}) = \beta \, \|\boldsymbol{x}\|_2^2$$




- Colors show equivalent (normalized) prior $p(\boldsymbol{x}) \,/\, p(\mathbf{0}) = e^{-R(\boldsymbol{x})}$
- Equivalent to IID gaussian prior on $\boldsymbol{x}$
- Makes any ill-conditioned / ill-posed problem well conditioned
- Ignores correlations between pixels

# Sparsity regularization in ambient space

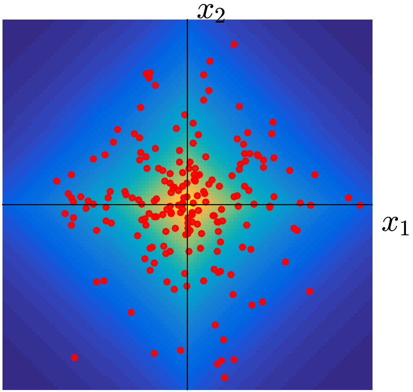
$$R(\boldsymbol{x}) = \beta \|\boldsymbol{x}\|_0 = \beta \sum_j \mathbb{I}_{\{x_j \neq 0\}}$$



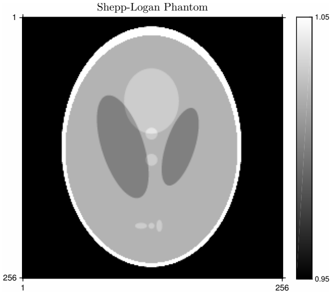

- Approximate Bayesian interpretation
- Non-convex
- IID $\Longrightarrow$ also ignores correlations

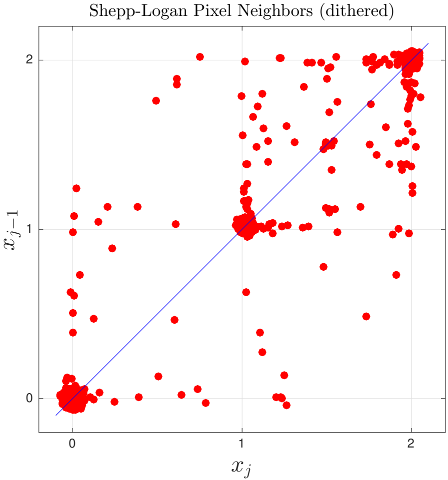
# Sparsity regularization: convex relaxation

$$\mathsf{R}(\boldsymbol{x}) = \beta \, \|\boldsymbol{x}\|_1 = \beta \sum_j |x_j|$$




- Equivalent to IID Laplacian prior on $\boldsymbol{x}$
- Also ignores correlations



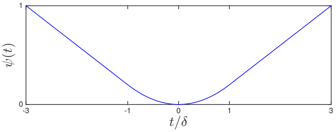

Caution: Shepp-Logan phantom [3] was designed for testing non-Bayesian methods, not for designing signal models. Q: What causes the spread??

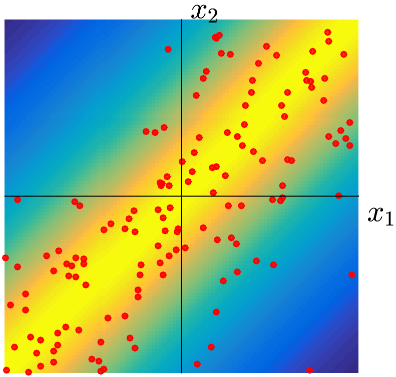
# Edge-preserving regularization

Neighboring pixels tend to have similar values except near edges:

$$\mathsf{R}(\boldsymbol{x}) = \beta \sum_j \psi(x_j - x_{j-1})$$

Potential function $\psi$:



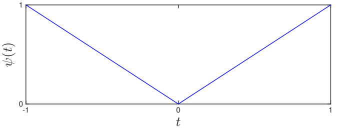

- Equivalent to improper prior (agnostic to DC value)
- Accounts for spatial correlations, but only very locally
- Used clinically now for low-dose X-ray CT image reconstruction

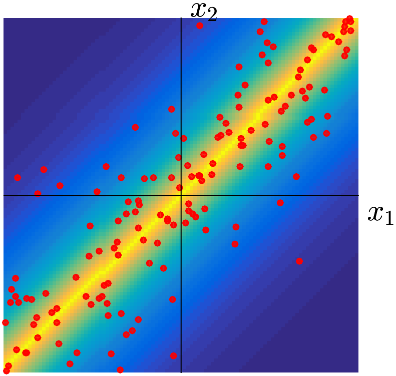
# Total-variation (TV) regularization

Neighboring pixels tend to have similar values except near edges ("gradient sparsity"):

$$\mathrm{R}(\boldsymbol{x}) = \beta \mathrm{TV}(\boldsymbol{x}) = \beta \, \|\boldsymbol{C}\boldsymbol{x}\|_1$$
$$= \beta \sum_j |x_j - x_{j-1}|$$

Potential function $\psi$:




- Equivalent to improper prior (agnostic to DC value)
- Accounts for correlations, but only very locally
- Well-suited to piece-wise constant Shepp-Logan phantom!
- Used in many academic publications...

# Black-box denoiser as a regularizer

$$\text{Noisy image} \rightarrow \boxed{\text{Denoiser}} \rightarrow \text{Denoised image}$$

- Example: Non-local means (NLM)
- Corresponding regularizer [4]–[6]:

$$\mathsf{R}(\boldsymbol{x}) = \beta \frac{1}{2} \|\boldsymbol{x} - \mathrm{NLM}(\boldsymbol{x})\|_2^2$$

- Encourages self-consistency with denoised version of image
- No evident Bayesian interpretation
- Variable splitting can facilitate minimization [7].

# Many more regularizers / priors

- Transforms: wavelets, curvelets, ...
- Markov random field models
- Graphical models
- ...

# Outline

Why

- Low-dose X-ray CT imaging
- Accelerated MR imaging
- Other inverse problems

How

- MAP estimation for inverse problems
- Classical regularization methods (some sparsity based)
- Contemporary regularization methods (all sparsity based)

# Contemporary regularizers

- Convolutional sparsity
- Union of subspaces
- Sparse coding with dictionary
- manifolds? [8]

# Convolutional sparsity

Idea:

$$x[\vec{n}] \approx \sum_{k=1}^{K} h_k[\vec{n}] * z_k[\vec{n}]$$

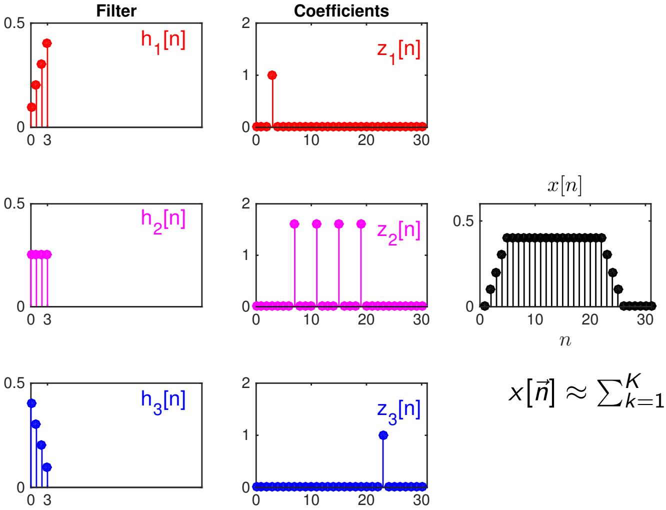
- where each $h_k[\vec{n}]$ is a FIR filter with $\|\boldsymbol{h}_k\| = 1$
- and each coefficient image $z_k[\vec{n}]$ is sparse [9]–[11].

Equivalent matrix-vector representation:

$$\boldsymbol{x} \approx \sum_{k=1}^{K} \boldsymbol{H}_k \boldsymbol{z}_k$$

where $\boldsymbol{H}_k$ is a Toeplitz (or circulant) matrix corresponding to $\boldsymbol{h}_k$.

# Convolutional sparsity: example




$$x[\vec{n}] \approx \sum_{k=1}^{K} h_k[\vec{n}] * z_k[\vec{n}]$$

# Convolutional sparsity: regularizer

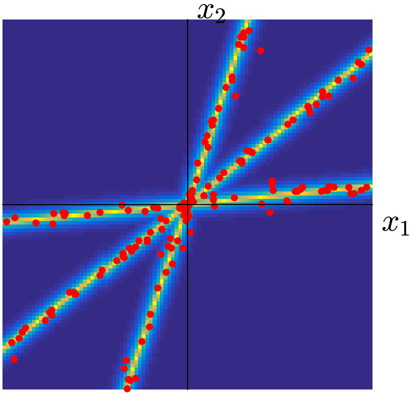
Recall $\boldsymbol{x} \approx \sum_{k=1}^{K} \boldsymbol{H}_k \boldsymbol{z}_k$
Natural corresponding regularizer:

$$\mathsf{R}(\boldsymbol{x}) = \min_{\{\boldsymbol{z}_k\}} \min_{\substack{\{\boldsymbol{h}_k\} \\ \|\boldsymbol{h}_k\| = 1}} \left( \left\| \boldsymbol{x} - \sum_{k=1}^{K} \boldsymbol{H}_k \boldsymbol{z}_k \right\|_2^2 + \lambda^2 \sum_{k=1}^{K} \|\boldsymbol{z}_k\|_0 \right)$$

Adapts FIR filters $\{\boldsymbol{h}_k\}$ and coefficients $\{\boldsymbol{z}_k\}$ to candidate $\boldsymbol{x}$.

- Literature focuses on the minimization problem (sparse coding)
- Yet to be explored as regularizer for inverse problems
- Inherently shift-invariant representation; no "patches" needed

# Union of subspaces model




- Dimensionality reduction?
- *cf.* classification / clustering motivation [12]
- (Extension to union of "flats" (linear varieties) is possible [13].)

# Union of subspaces regularization

Given (?) collection of $K$ subspace bases $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_K$ (dictionaries with full column rank):

$$\mathsf{R}(\boldsymbol{x}) = \underbrace{\min_{k}}_{\text{“classification”}} \underbrace{\min_{\boldsymbol{z}_k} \quad \beta \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{D}_k \boldsymbol{z}_k\|_2^2}_{\text{regression}}$$

$$= \min_{k} \beta \frac{1}{2} \left\| \boldsymbol{x} - \boldsymbol{D}_k \boldsymbol{D}_k^{+} \boldsymbol{x} \right\|_2^2$$

- $\mathsf{R}(\boldsymbol{x}) = 0$ if $\boldsymbol{x}$ lies in the span of any of the dictionaries $\{\boldsymbol{D}_k\}$.
- otherwise, distance to nearest subspace (discourage, not constrain)
- Non-convex (highly?) (*cf.* preceding picture)
- Apply to image patches to be practical
- Equivalent Bayesian interpretation? (not a mixture model here)

# Regularizers using sparse coding with dictionaries

Assume $\boldsymbol{x} \approx \boldsymbol{D}\boldsymbol{z}$ where $\boldsymbol{D}$ is a dictionary (often over-complete) and $\boldsymbol{z}$ is a sparse coefficient vector. Corresponding regularizers:

$$\mathsf{R}(\boldsymbol{x}) = \min_{\boldsymbol{z} \,:\, \|\boldsymbol{z}\|_p \leq s} \beta \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{z}\|_2^2$$

$$\mathsf{R}(\boldsymbol{x}) = \min_{\boldsymbol{z}} \left( \beta_1 \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{z}\|_2^2 + \beta_2 \|\boldsymbol{z}\|_p \right)$$

- Convex in $\boldsymbol{z}$ (for given $\boldsymbol{x}$) if $p \geq 1$.
- $\mathsf{R}(\boldsymbol{x})$ typically non-convex in $\boldsymbol{x}$.
- Could be equivalent to a union-of-subspaces regularizer if $\boldsymbol{D} = [\boldsymbol{D}_1 \ \ldots \ \boldsymbol{D}_K]$ and if we constrain coefficient vector $\boldsymbol{z}$ in a non-standard way.

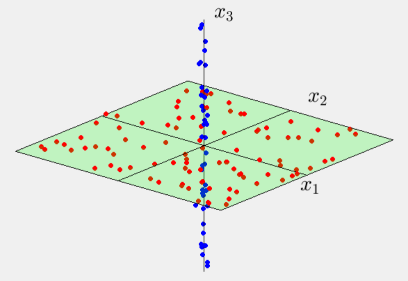
# Union-of-subspaces vs sparse-coding-with-dictionary

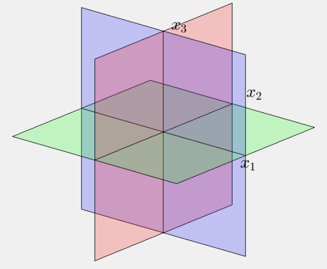
Consider union-of-subspaces model with $\boldsymbol{D}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$, $\boldsymbol{D}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$.

So $\boldsymbol{D}_1$ spans x-y plane and $\boldsymbol{D}_2$ spans z-axis.

A dictionary model with $\boldsymbol{D} = [\boldsymbol{D}_1 \ \boldsymbol{D}_2] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

and sparsity $s = 2$, happily represents all three cardinal planes




Thus dictionary model seems "less constrained" than union-of-subspaces model.

(Still, focus on sparse dictionary representation hereafter.)

# Dictionary learning from training data

- Given training data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$ (image patches)
- Assumed model: $\boldsymbol{x}_n \approx \boldsymbol{D}\boldsymbol{z}_n$
- unknown $d \times J$ dictionary $\boldsymbol{D} = [\boldsymbol{d}_1 \; \ldots \; \boldsymbol{d}_J]$
- coefficient vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N \in \mathbb{R}^J$ assumed "sparse"

K-SVD dictionary learning formulation [14]:

$$\boldsymbol{D}^* = \underset{\boldsymbol{D} \in \mathbb{R}^{d \times J}}{\arg\min} \sum_{n=1}^{N} \min_{\boldsymbol{z}_n \in \mathbb{R}^J} \|\boldsymbol{x}_n - \boldsymbol{D}\boldsymbol{z}_n\|_2 \quad \text{s.t.} \quad \begin{array}{c} \|\boldsymbol{d}_j\| = 1 \ \forall j \\ \|\boldsymbol{z}_n\|_0 \leq s \ \forall n \end{array}$$

$$= \underset{\boldsymbol{D} \in \mathbb{R}^{d \times J}}{\arg\min} \min_{\boldsymbol{Z} \in \mathbb{R}^{J \times N}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{Z}\|_F \quad \text{s.t.} \quad \begin{array}{c} \|\boldsymbol{d}_j\| = 1 \ \forall j \\ \|\boldsymbol{z}_n\|_0 \leq s \ \forall n \end{array}$$

$\boldsymbol{X} \triangleq [x_1 \; \ldots \; x_N]$, $\boldsymbol{Z} \triangleq [z_1 \; \ldots \; z_N]$
Computationally expensive and no convergence guarantees.
Inherently non-convex due to product of unknowns $\boldsymbol{D}\boldsymbol{Z}$.

# New dictionary learning method (SOUP-DIL)

Joint work with Sai Ravishankar and Raj Nadakuditi [15]–[18]

- Write sparse representation as Sum of OUter Products (SOUP):

$$\boldsymbol{X} \approx \boldsymbol{D}\boldsymbol{Z} = \boldsymbol{D}\boldsymbol{C}' = \sum_{j=1}^{J} \boldsymbol{d}_j \boldsymbol{c}_j'$$

  where $\boldsymbol{Z}' = \boldsymbol{C} = [\boldsymbol{c}_1 \ \ldots \ \boldsymbol{c}_J] \in \mathbb{R}^{N \times J}$ (coefficients for each atom)
- Replace individual atom sparsity constraint $\|\boldsymbol{z}_n\|_0 \leq s$ with aggregate sparsity regularizer: $\|\boldsymbol{Z}\|_0 = \|\boldsymbol{C}\|_0$.
  - Natural for DIctionary Learning (DIL) from training data
  - Unnatural for image compression using sparse coding

SOUP-DIL $\ell_0$ formulation:

$$\boldsymbol{D}^* = \mathop{\arg\min}_{\boldsymbol{D} \in \mathbb{R}^{d \times J}} \min_{\boldsymbol{C} \in \mathbb{R}^{N \times J}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{C}'\|_F^2 + \lambda^2 \|\boldsymbol{C}\|_0 \quad \text{s.t.} \quad \begin{array}{l} \|\boldsymbol{d}_j\|_2 = 1 \ \forall j \\ \|\boldsymbol{c}_j\|_\infty \leq L \ \forall j \end{array}$$

# SOUP-DIL algorithm

SOUP-DIL formulation:

$$\boldsymbol{D}^* = \underset{\boldsymbol{D}\in\mathbb{R}^{d\times J}}{\arg\min} \min_{\boldsymbol{C}\in\mathbb{R}^{N\times J}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{C}'\|_F^2 + \lambda^2 \|\boldsymbol{C}\|_0 \quad \text{s.t.} \quad \begin{array}{l} \|\boldsymbol{d}_j\|_2 = 1 \; \forall j \\ \|\boldsymbol{c}_j\|_\infty \leq L \; \forall j \end{array}$$

- Block coordinate descent (BCD) algorithm
  - Sparse coding step for $\boldsymbol{C}$
  - Dictionary update step for $\boldsymbol{D}$
- Very simple update rules (low compute cost)
- Monotone descent of $\Psi(\boldsymbol{D}, \boldsymbol{C})$
- Convergence theorem: for any given initialization $(\boldsymbol{D}^0, \boldsymbol{C}^0)$, all accumulation points of sequence $(\boldsymbol{D}, \boldsymbol{C})$
  - are critical points of cost $\Psi$ and
  - are equivalent (reach same cost function value $\Psi^*$).
  - Furthermore: $\left\{\left\|\boldsymbol{D}^{(k)} - \boldsymbol{D}^{(k-1)}\right\|\right\} \to 0$. Same for $\left\{\boldsymbol{C}^{(k)}\right\}$.

# SOUP-DIL updates

$$\boldsymbol{D}^* = \arg\min_{\boldsymbol{D}\in\mathbb{R}^{d\times J}} \min_{\boldsymbol{C}\in\mathbb{R}^{N\times J}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{C}'\|_F^2 + \lambda^2 \|\boldsymbol{C}\|_0 \quad \text{s.t.} \quad \begin{matrix} \|\boldsymbol{d}_j\|_2 = 1 \ \forall j \\ \|\boldsymbol{c}_j\|_\infty \leq L \ \forall j \end{matrix}$$

Alternate: update one column $\boldsymbol{d}_j$ of $\boldsymbol{D}$ then one column $\boldsymbol{c}_j$ of $\boldsymbol{C}$.

- Sparse coding step: update $\boldsymbol{c}_j$ using residual $\boldsymbol{E}_j \triangleq \sum_{k\neq j} \boldsymbol{d}_k \boldsymbol{c}_k'$

$$\min_{\boldsymbol{c}_j} \|\boldsymbol{E}_j - \boldsymbol{d}_j \boldsymbol{c}_j'\|_F^2 + \lambda^2 \|\boldsymbol{c}_j\|_0 \quad \text{s.t.} \quad \|\boldsymbol{c}_j\|_\infty \leq L$$

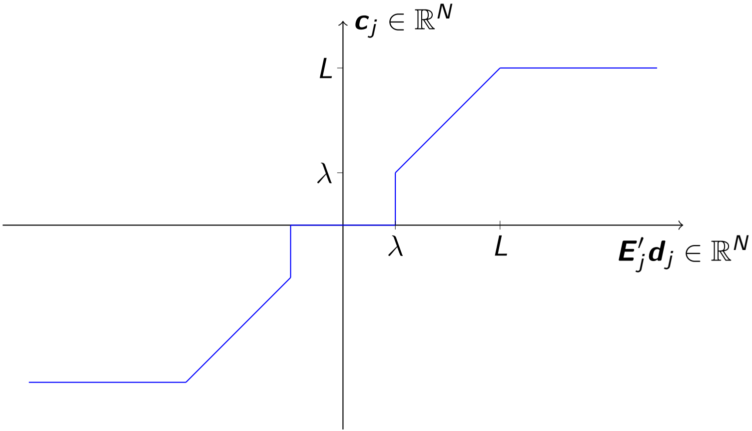
  Truncated (via $L$) hard thresholding of $\boldsymbol{E}_j'\boldsymbol{d}_j$ with threshold $\lambda$

- Dictionary atom step: update $\boldsymbol{d}_j$

$$\min_{\boldsymbol{d}_j} \|\boldsymbol{E}_j - \boldsymbol{d}_j \boldsymbol{c}_j'\|_F^2 \quad \text{s.t.} \quad \|\boldsymbol{d}_j\|_2 = 1$$

  Constrained least-squares solution: $\boldsymbol{d}_j = (\boldsymbol{E}_j \boldsymbol{c}_j) / \|\boldsymbol{E}_j \boldsymbol{c}_j\|_2$

# Truncated hard thresholding for SOUP-DIL

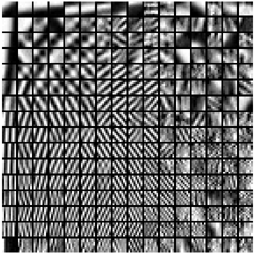



(Acts element-wise.) (In practice take $L = \infty$.)

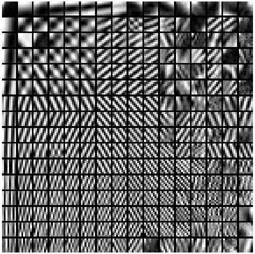
(Algorithm also provides a simple sparse coding method.)

# Example: dictionary learning for Barbara

Barbara



K-SVD $\boldsymbol{D}$



SOUP-DIL $\boldsymbol{D}$

Denoising PSNR (dB) from [15]

| $\sigma$ | Noisy | O-DCT | K-SVD | SOUP-DIL |
|---|---|---|---|---|
| 20 | 22.13 | 29.95 | **30.83** | 30.79 |
| 25 | 20.17 | 28.68 | 29.63 | **29.64** |
| 30 | 18.59 | 27.62 | 28.54 | **28.63** |
| 100 | 8.11 | 21.87 | 21.87 | **21.97** |

SOUP-DIL faster than K-SVD

# Regularization using SOUP-DIL

- Large image $\boldsymbol{x}$, extract $M$ patches $\boldsymbol{X} = [\boldsymbol{P}_1\boldsymbol{x} \ \ldots \ \boldsymbol{P}_M\boldsymbol{x}]$
- Assume patch $\boldsymbol{x}_m = \boldsymbol{P}_m\boldsymbol{x} \approx \boldsymbol{D}\boldsymbol{z}_m$ has (aggregate) sparse representation in dictionary $\boldsymbol{D} \in \mathbb{R}^{d \times J}$ where $d$ is patch size

$$\mathrm{R}(\boldsymbol{x}) = \mathrm{R}(\boldsymbol{X}) = \min_{\boldsymbol{C} \in \mathbb{R}^{M \times J}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{C}'\|_F^2 + \lambda^2 \|\boldsymbol{C}\|_0 \quad \text{s.t. } \|\boldsymbol{c}_j\|_\infty \leq L \ \forall j$$

- $\mathrm{R}(\boldsymbol{x}) = 0$ if patches can be represented exactly with "sufficiently few" non-zero coefficients (depends on $\lambda$)
- Ignore constraint $\|\boldsymbol{c}_j\|_\infty \leq L$
- Bayesian interpretation?

# CT reconstruction using (known) dictionary regularizer

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{W}}^2 + \beta\, \mathrm{R}(\boldsymbol{x})$$
$$= \arg\min_{\boldsymbol{x}} \min_{\boldsymbol{C} \in \mathbb{R}^{M \times J}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{W}}^2 + \frac{\beta}{2} \left( |\!|\!| \boldsymbol{X} - \boldsymbol{D}\boldsymbol{C}' |\!|\!|_F^2 + \lambda^2 |\!|\!| \boldsymbol{C} |\!|\!|_0 \right)$$

Alternating (nested) minimization:

- Fixing $\boldsymbol{x}$, updating each column of $\boldsymbol{C}$ sequentially involves (truncated?) hard-thresholding
- Fixing $\boldsymbol{C}$, updating $\boldsymbol{x}$ is (large-scale) quadratic problem

$$g(\boldsymbol{x}) = \frac{1}{2} |\!|\!| \boldsymbol{X} - \boldsymbol{D}\boldsymbol{C}' |\!|\!|_F^2 = \sum_{m=1}^{M} \frac{1}{2} \left\| \boldsymbol{P}_m \boldsymbol{x} - \boldsymbol{D} \boldsymbol{P}_m \boldsymbol{C}' \right\|_2^2$$

$$\nabla^2 g(\boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{P}_m' \boldsymbol{P}_m \quad \text{is diagonal}$$

- Work in progress...

# Why Bayesian?

- Numerous "normal-dose" CT images!
- learn $\boldsymbol{D}$, or most of it, from "big data"
- learn statistics of sparse coefficients $\boldsymbol{Z}$?
- replace generic $\|\boldsymbol{Z}\|_0$ with $\mathrm{p}(\boldsymbol{Z})$?

# Extensions / future work

- Use majorization to update multiple columns of $\boldsymbol{D}$ or $\boldsymbol{C}$ simultaneously
- DC atom
- Rotate/flip atoms [19] [20]
- rank constraints on dictionary atoms [16]
- Tensor structured atoms for 3D / dynamic imaging
- Combined transform learning / dictionary learning
- Union of manifolds instead of union of subspaces?
- ...

Open problems

- Model selection
- Parameter selection
- Performance guarantees

[1] S. Ravishankar and Y. Bresler, “MR image reconstruction from highly undersampled k-space data by dictionary learning,” *IEEE Trans. Med. Imag.*, vol. 30, no. 5, 1028–41, May 2011.

[2] E. Candès and J. K. Romberg, “Signal recovery from random projections,” in *Proc. SPIE 5674 Computational Imaging III*, 2005, 76–86.

[3] L. A. Shepp and B. F. Logan, “The Fourier reconstruction of a head section,” *IEEE Trans. Nuc. Sci.*, vol. 21, no. 3, 21–43, Jun. 1974.

[4] M. Mignotte, “A non-local regularization strategy for image deconvolution,” *Pattern Recognition Letters*, vol. 29, no. 16, 2206–12, Dec. 2008.

[5] Z. Yang and M. Jacob, “Nonlocal regularization of inverse problems: A unified variational framework,” *IEEE Trans. Im. Proc.*, vol. 22, no. 8, 3192–203, Aug. 2013.

[6] H. Zhang, J. Ma, J. Wang, Y. Liu, H. Lu, and Z. Liang, “Statistical image reconstruction for low-dose CT using nonlocal means-based regularization,” *Computerized Medical Imaging and Graphics*, vol. 38, no. 6, 423–35, Sep. 2014.

[7] S. Y. Chun, Y. K. Dewaraja, and J. A. Fessler, “Alternating direction method of multiplier for tomography with non-local regularizers,” *IEEE Trans. Med. Imag.*, vol. 33, no. 10, 1960–8, Oct. 2014.

[8] S. Poddar and M. Jacob, “Dynamic MRI using smooThness regularization on manifolds (SToRM),” *IEEE Trans. Med. Imag.*, 2016.

[9] J. Yang, K. Yu, and T. Huang, “Supervised translation-invariant sparse coding,” in *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition*, 2010, 3517–24.

[10] H. Bristow, A. Eriksson, and S. Lucey, “Fast convolutional sparse coding,” in *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition*, 2013, 391–8.

[11] B. Wohlberg, “Efficient convolutional sparse coding,” in *Proc. IEEE Conf. Acoust. Speech Sig. Proc.*, 2014, 7173–7.

# Bibliography II


[12] R. Vidal, "Subspace clustering," *IEEE Sig. Proc. Mag.*, vol. 28, no. 2, 52–68, Mar. 2011.

[13] T. Zhang, A. Szlam, and G. Lerman, "Median K-Flats for hybrid linear modeling with many outliers," in *Proc. Intl. Conf. Comp. Vision*, 2009, 234–41.

[14] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Sig. Proc.*, vol. 54, no. 11, 4311–22, Nov. 2006.

[15] S. Ravishankar, R. R. Nadakuditi, and J. A. Fessler, "Efficient sum of outer products dictionary learning (SOUP-DIL) - The $\ell_0$ method," *IEEE Trans. Sig. Proc.*, 2015, Submitted.

[16] ——,"Efficient learning of dictionaries with low-rank atoms," in *Proc. IEEE Intl. Conf. on Image Processing*, Submitted., 2016.

[17] ——,"Efficient l0 dictionary learning with convergence guarantees," in *Proc. Intl. Conf. Mach. Learn*, Submitted., 2016.

[18] ——,*Efficient sum of outer products dictionary learning (SOUP-DIL) - The $\ell_0$ method*, arxiv 1511.08842, 2015.

[19] B. Wen, S. Ravishankar, and Y. Bresler, *FRIST - flipping and rotation invariant sparsifying transform learning and applications of inverse problems*, 2016.

[20] B. Wen, Y. Bresler, and S. Ravishankar, *FRIST - flipping and rotation invariant sparsifying transform learning and applications*, arxiv 1511.06359, 2015.