SINBAD Consortium 2010

Algorithms for sparse optimization

Michael P. Friedlander

SLIM
University of British Columbia

Outline

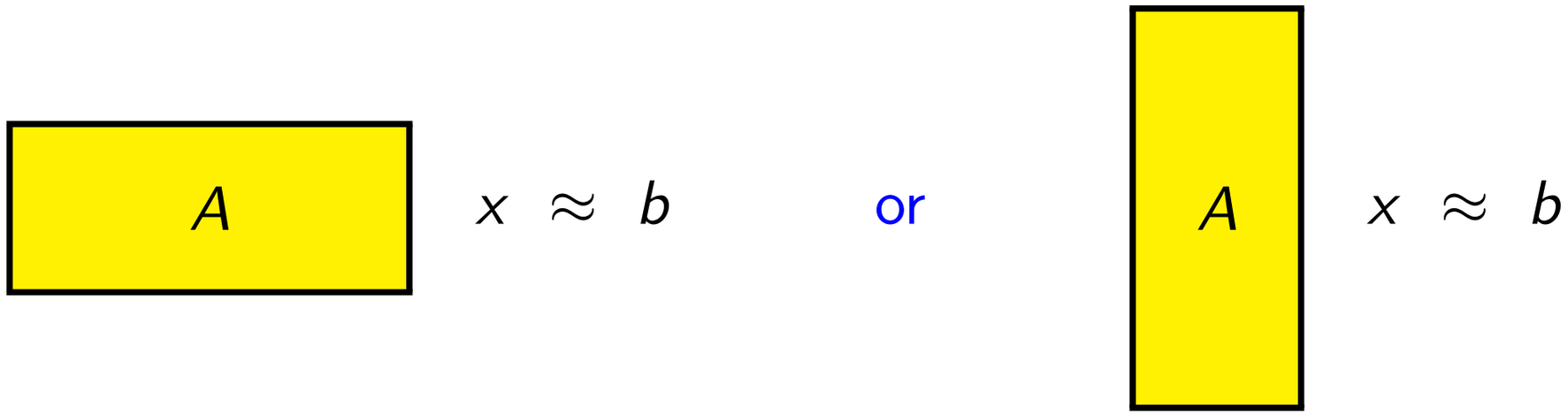
- motivation
- example applications and formulations
- main algorithmic approaches

  projected gradient, iterative soft-thresholding, proximal point, augmented Lagrangian, Bregman, alternating direction of method of multipliers, split Bregman

- sample research topics
- software

MOTIVATION

Sparse solutions



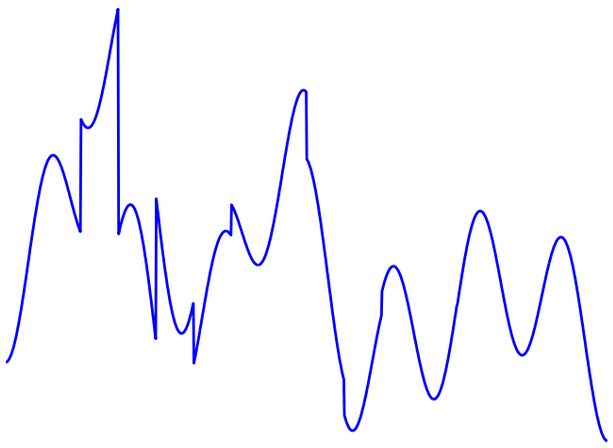
Problem: find sparse solution x

$$\text{minimize} \quad \text{nnz}(x) \quad \text{subj to} \quad Ax \approx b$$

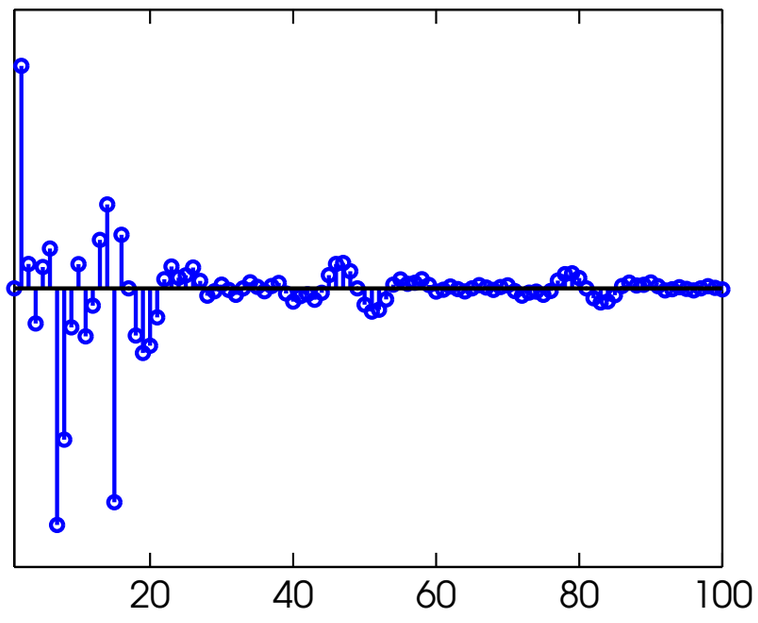
Drawbacks

- combinatorial problem
- tractable for only trivial problems

Sparse representations

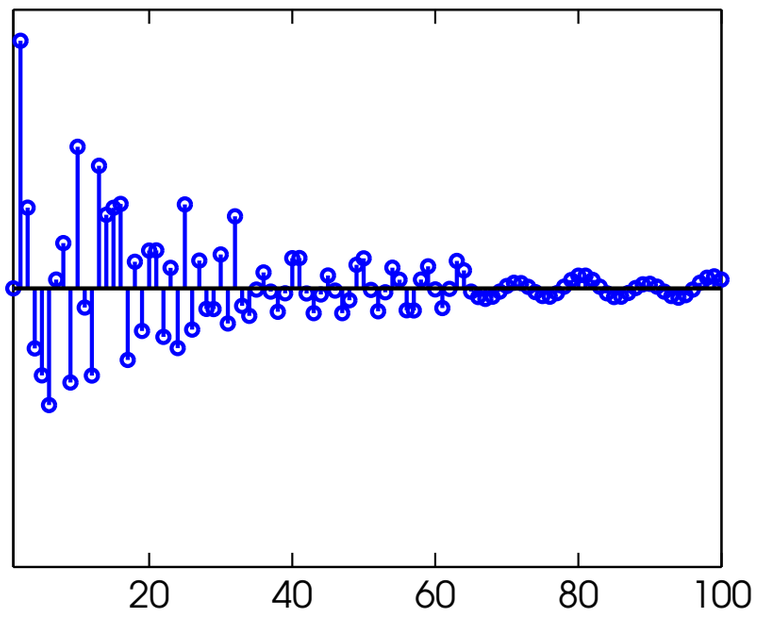


$y = Hx$



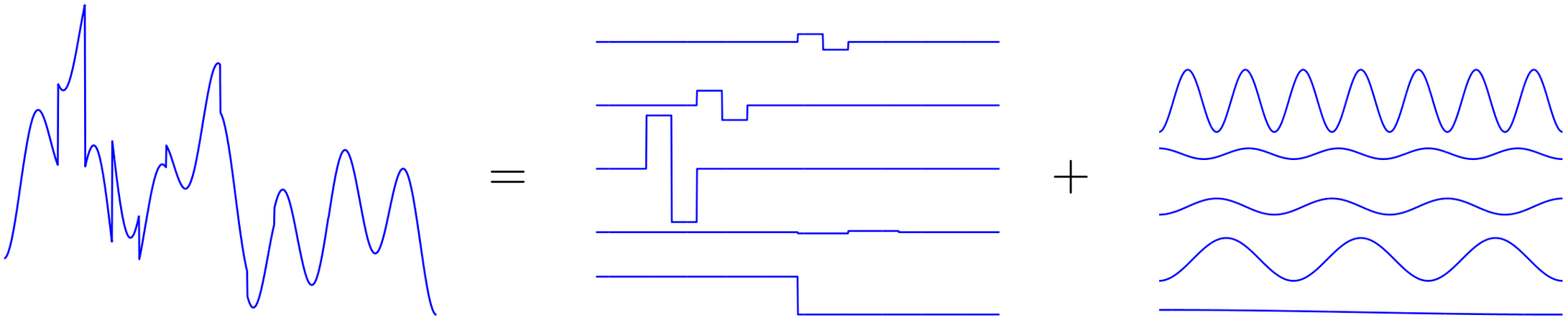
Haar

$y = Dx$

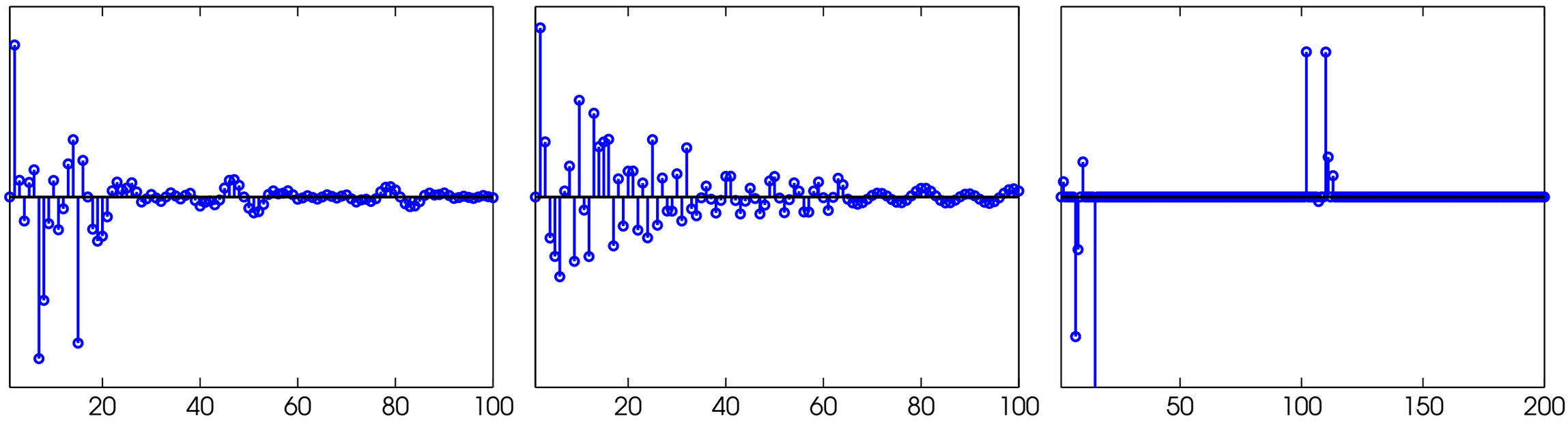


DCT

Sparse representations



$y = Hx$ $\qquad$ $y = Dx$ $\qquad$ $y = \begin{bmatrix} H & D \end{bmatrix} x$



Haar $\qquad$ DCT $\qquad$ Haar/DCT dictionary

Sparsity via a transform

- represent a signal y as a superposition of elementary signal atoms:

$$y = \sum_j \phi_j x_j \quad \text{ie,} \quad y = \Phi x$$

Sparsity via a transform

- represent a signal y as a superposition of elementary signal atoms:

$$y = \sum_j \phi_j x_j \quad \text{ie,} \quad y = \Phi x$$

- orthogonal bases Φ have unique representation $x = \Phi^T y$

Sparsity via a transform

- represent a signal y as a superposition of elementary signal atoms:

$$y = \sum_j \phi_j x_j \quad \text{ie,} \quad y = \Phi x$$

- orthogonal bases Φ have unique representation $x = \Phi^T y$
- overcomplete dictionaries, eg, $A = [\Phi_1 \ \Phi_2]$ have more flexibility, but representation is not unique. One approach:

$$\text{minimize} \ \ \text{nnz}(x) \ \ \text{subj to} \ \ Ax \approx y$$

Sparsity via a transform

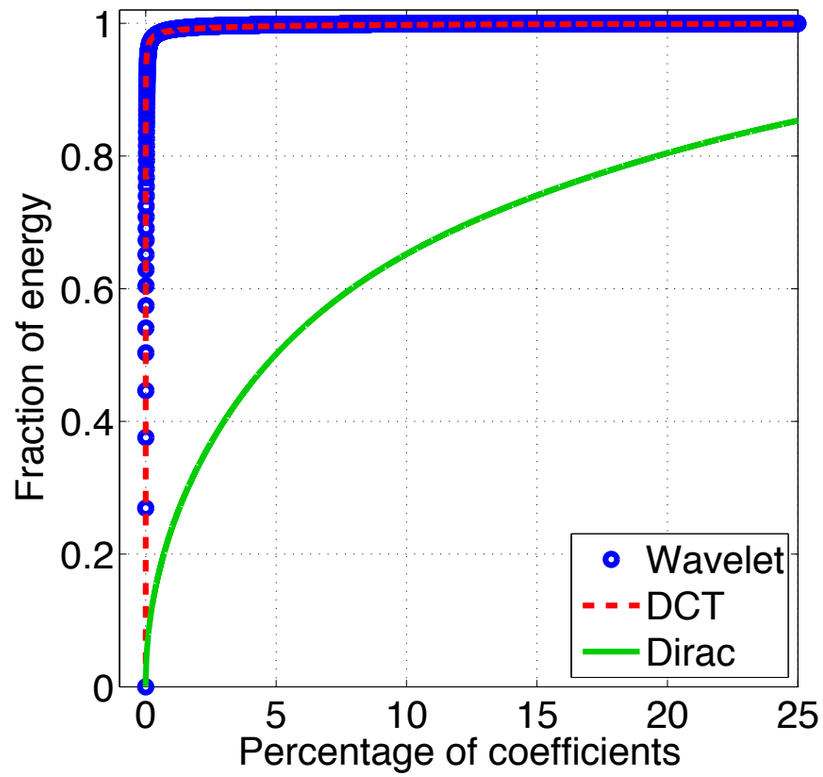
- represent a signal y as a superposition of elementary signal atoms:

$$y = \sum_j \phi_j x_j \quad \text{ie,} \quad y = \Phi x$$

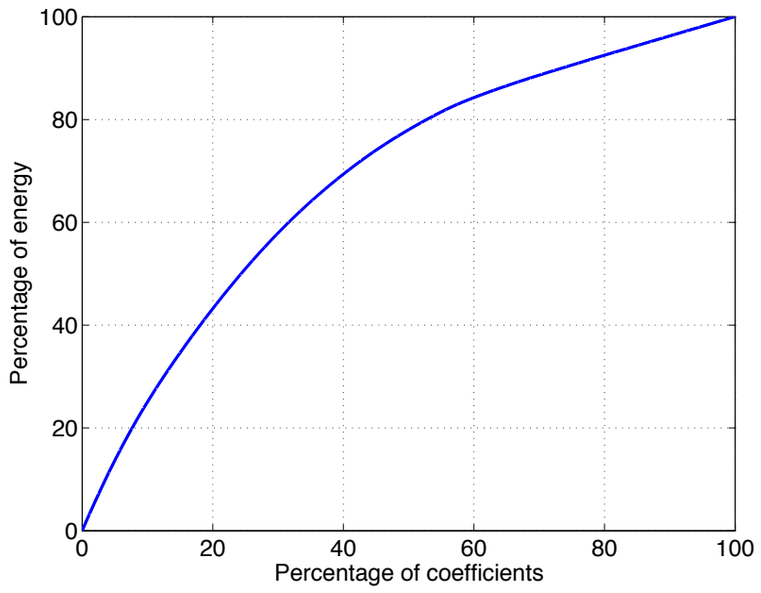
- orthogonal bases Φ have unique representation $x = \Phi^T y$
- overcomplete dictionaries, eg, $A = [\Phi_1\ \Phi_2]$ have more flexibility, but representation is not unique. One approach:

$$\text{minimize } \text{nnz}(x) \text{ subj to } Ax \approx y$$

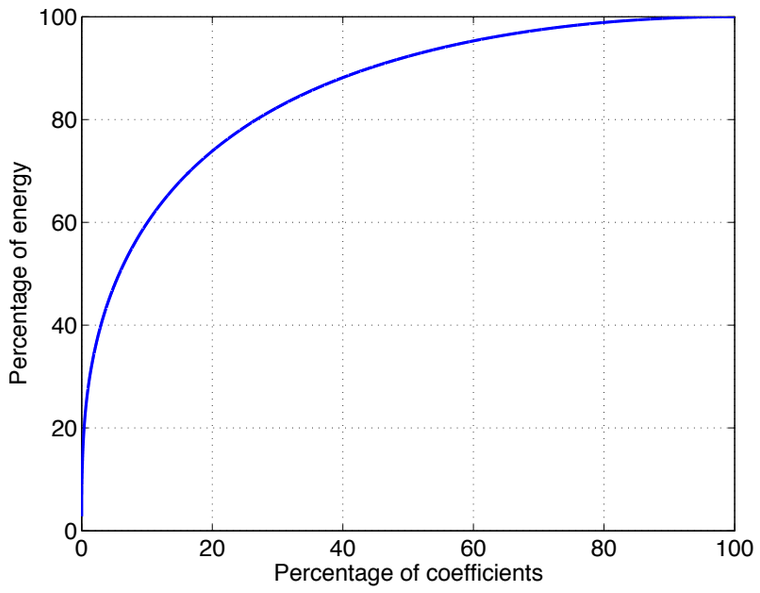
- "efficient" transforms lead to fast decay in coefficients x_j



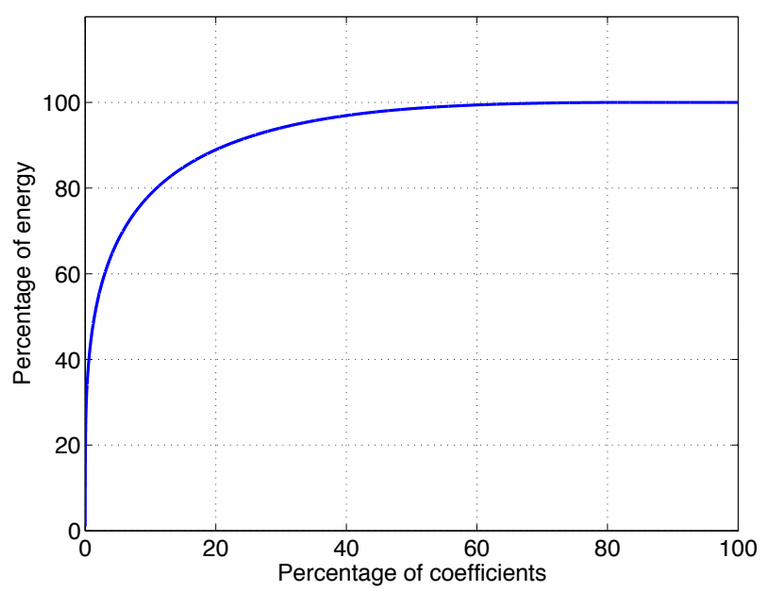
Approximate sparsity



Identity

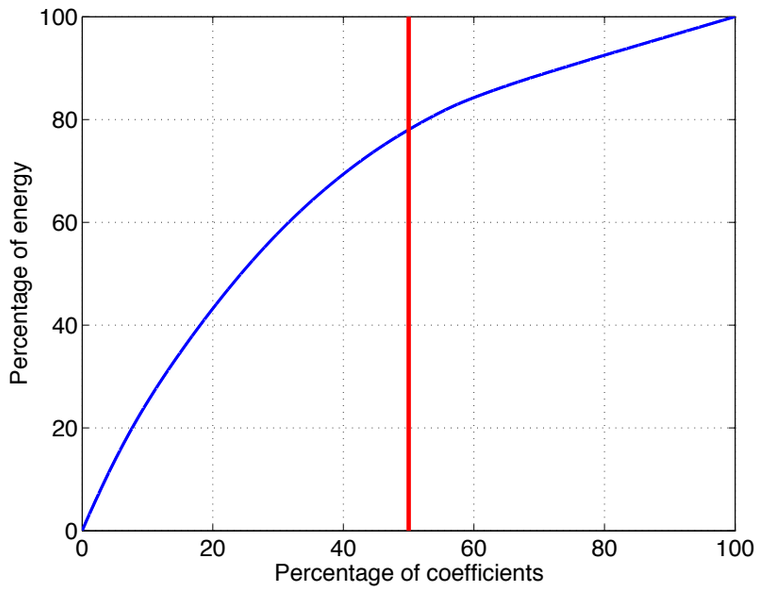


Fourier

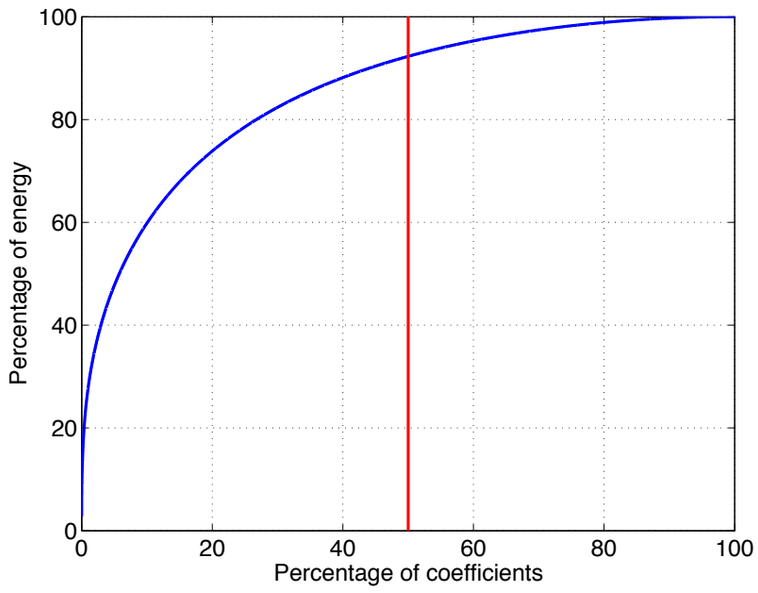


Wavelet

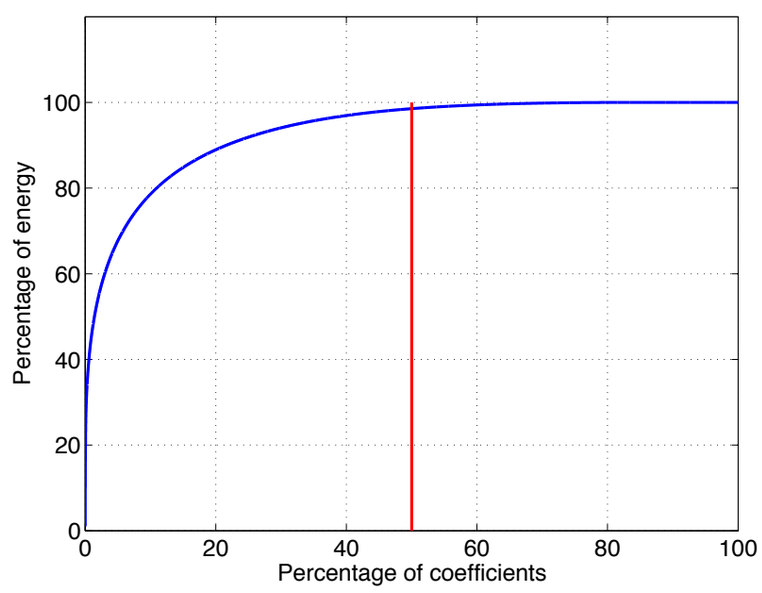
Approximate sparsity



Identity

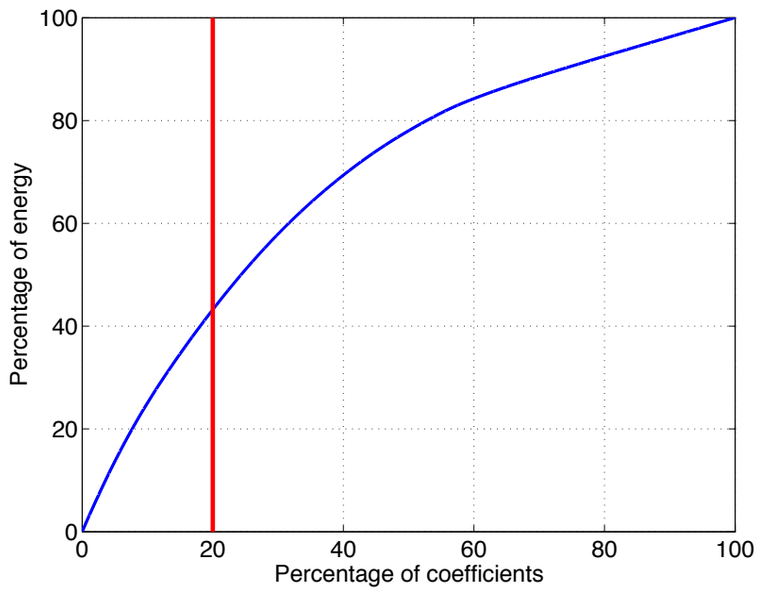


Fourier

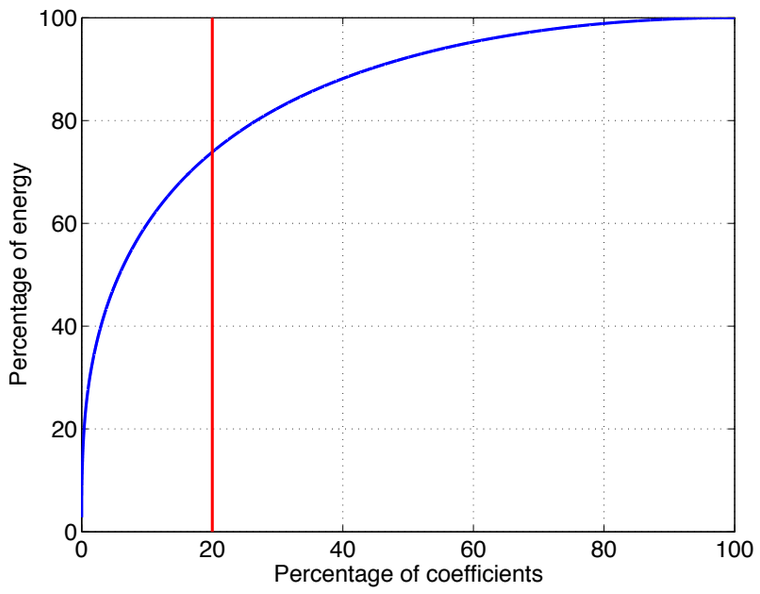


Wavelet

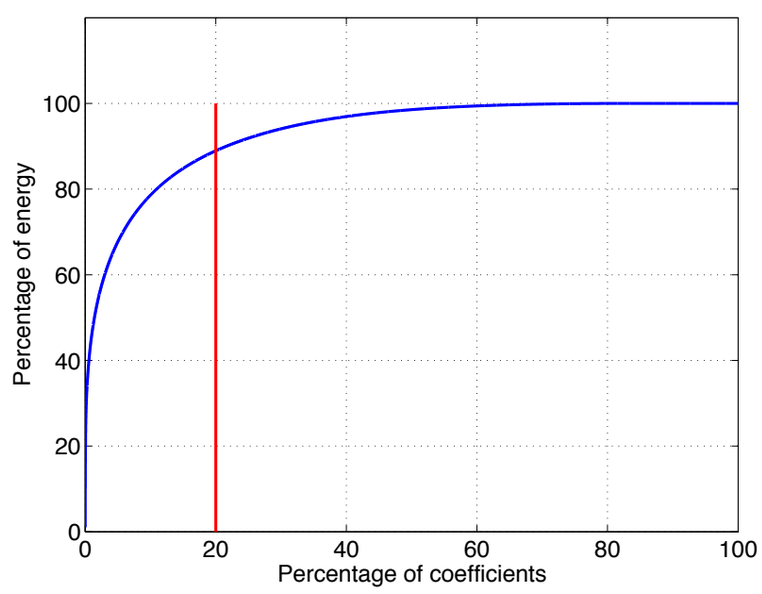
Approximate sparsity



Identity

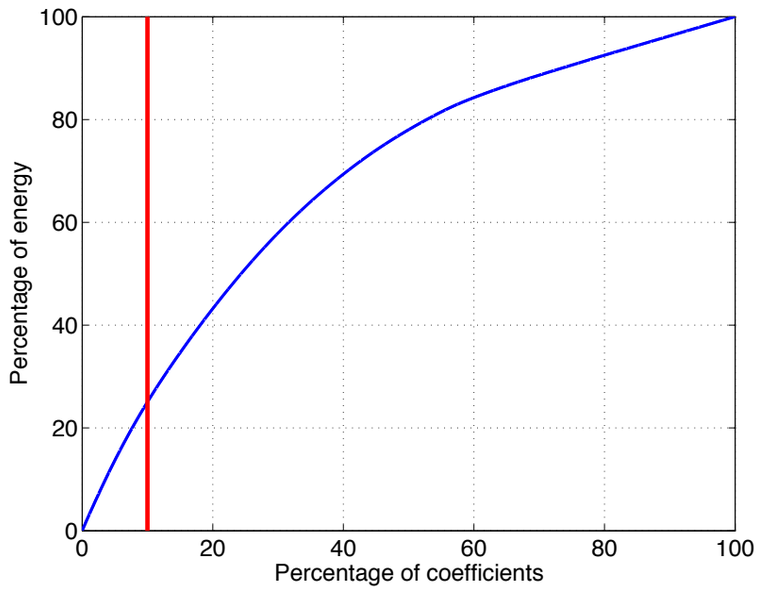


Fourier

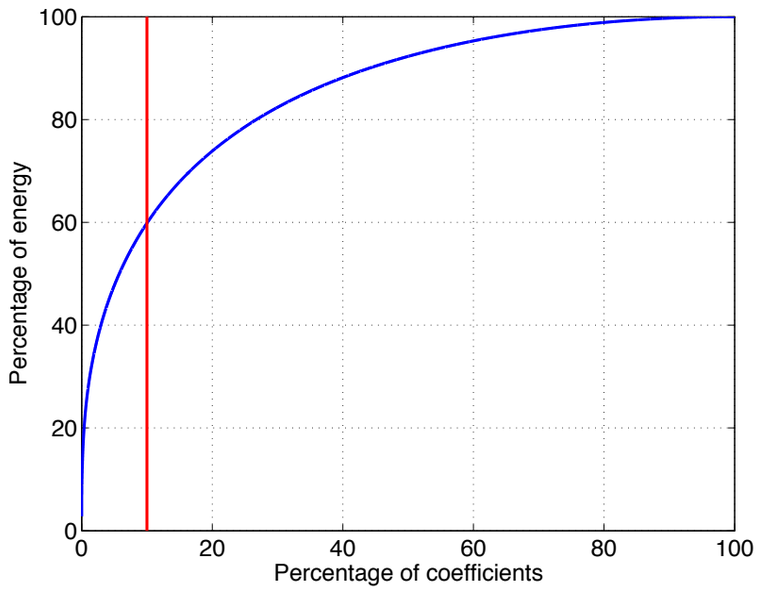


Wavelet

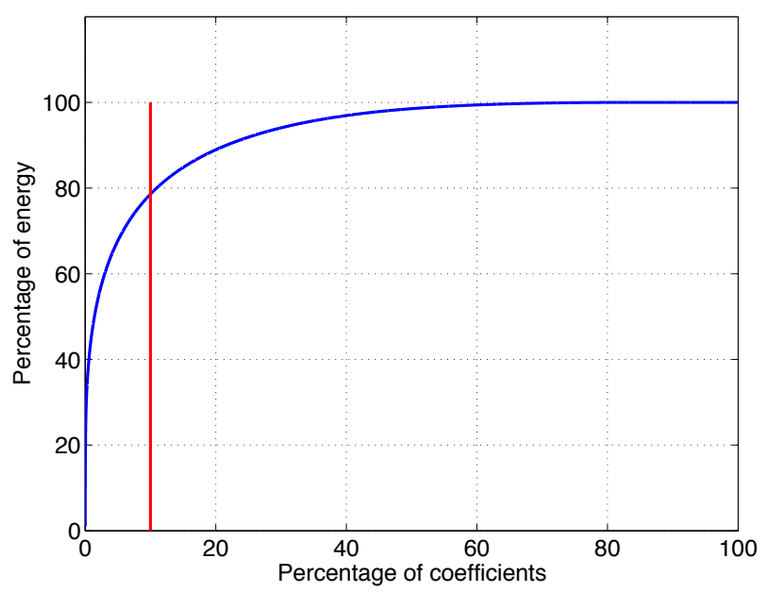
Approximate sparsity



Identity

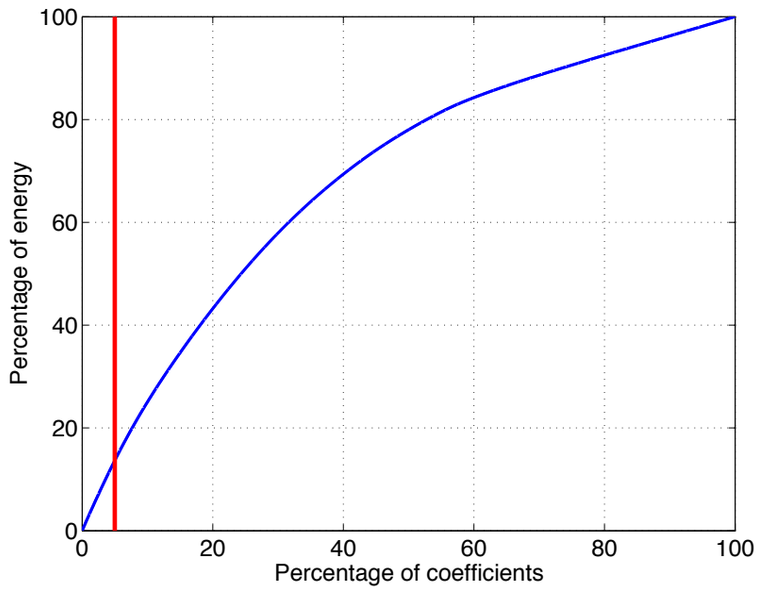


Fourier

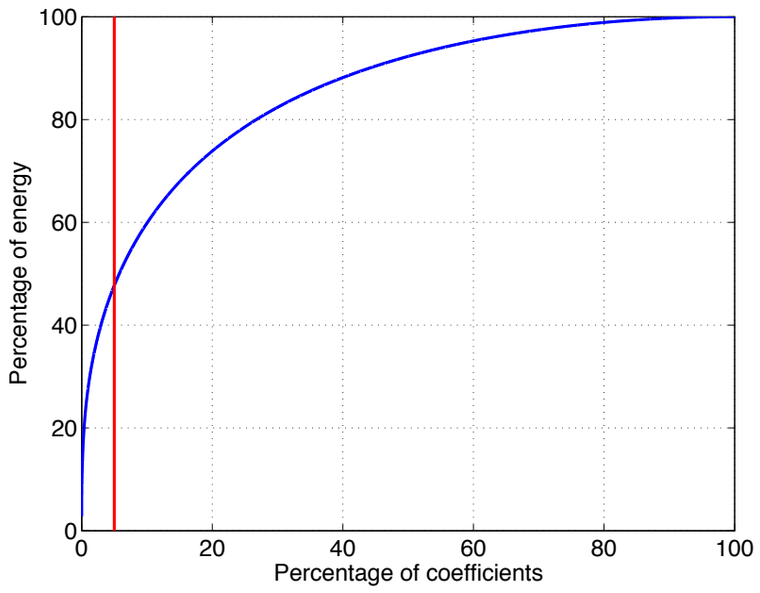


Wavelet

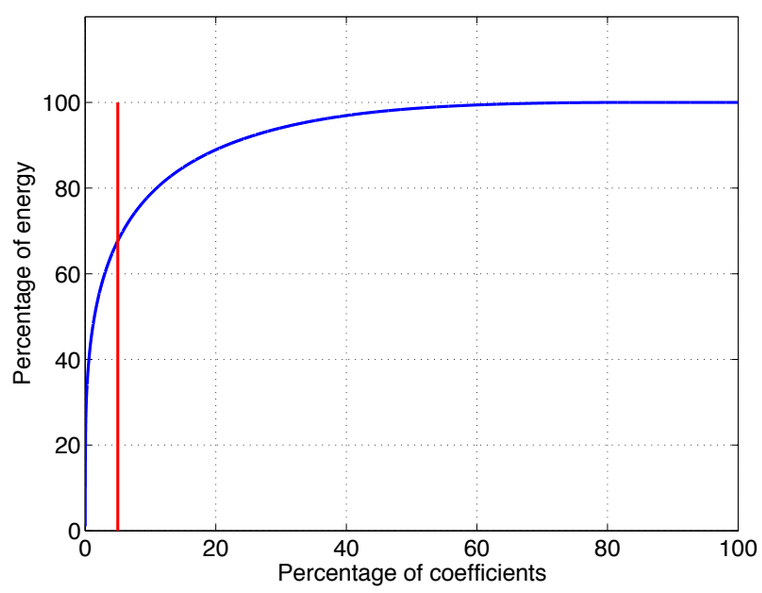
Approximate sparsity



Identity

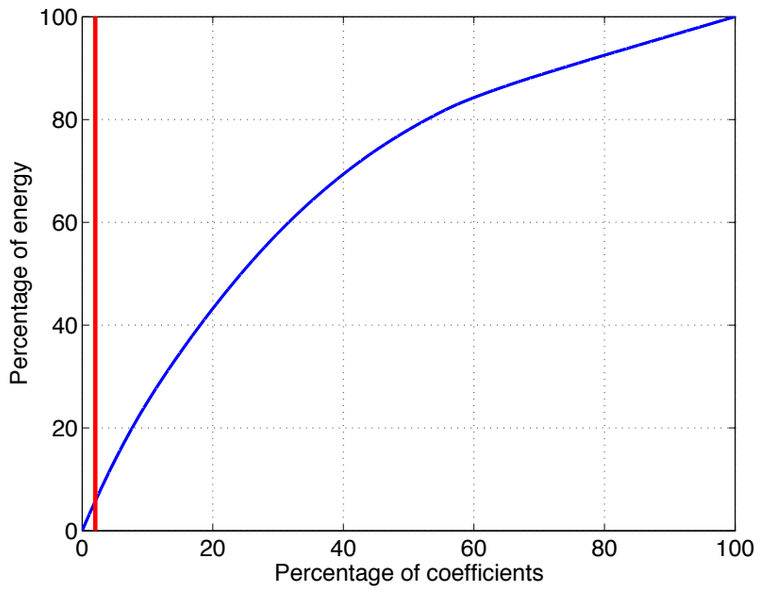


Fourier

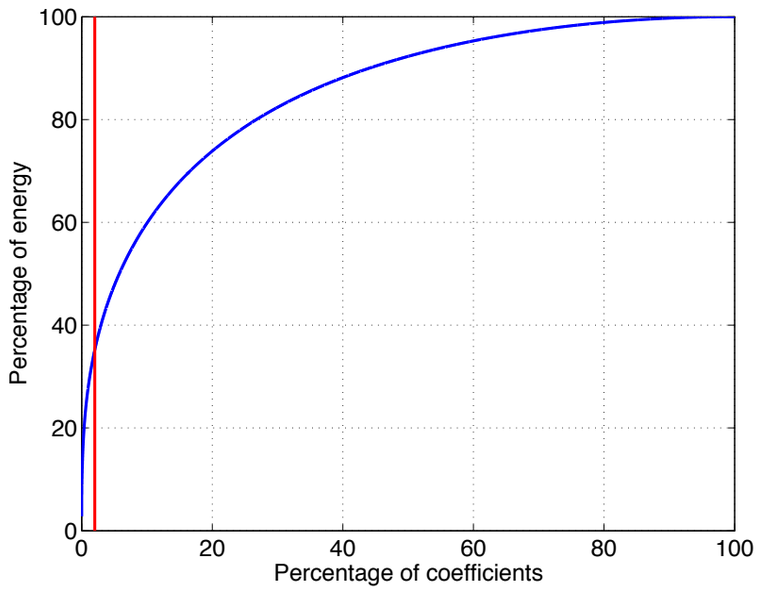


Wavelet

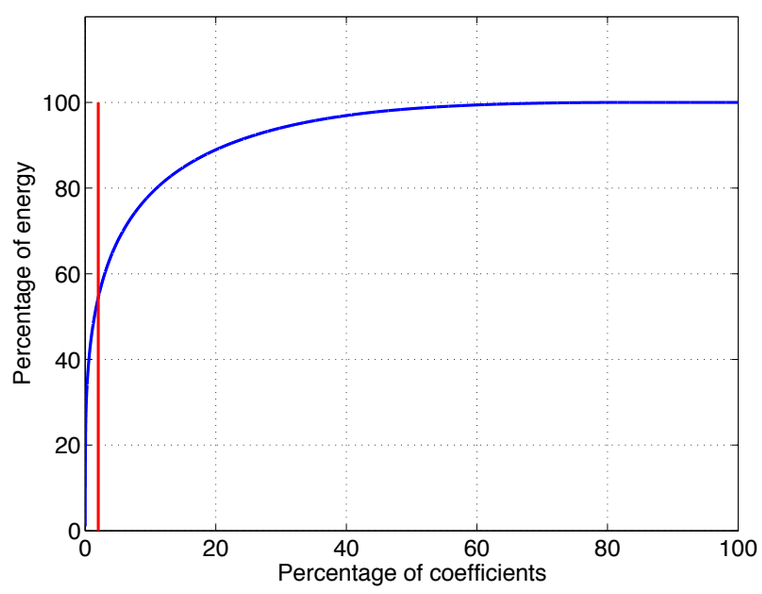
Approximate sparsity



Identity



Fourier



Wavelet

Approaches

Greedy algorithms choose nonzero x_j

- *one by one:* OMP $\quad \min \ \|S_k x - b\|_2, \quad S_{k+1} \leftarrow \begin{bmatrix} S_k & a_j \end{bmatrix}$
- *batches:* CoSaMP $\quad \min \ \|S_k x - b\|_2, \quad S_{k+1} \leftarrow \begin{bmatrix} A_{\{s \text{ elements}\}} \end{bmatrix}$

[Tropp-Gilbert 07; Tropp-Needell 08]

Convex relaxations

- minimize $\quad \|x\|_1 \quad$ subj to $\quad Ax \approx b$

[Donoho 05; Candès-Tao 06; Candès 06]

Nonconvex heuristics

- minimize $\quad \|x\|_p \quad$ subj to $\quad Ax \approx b, \quad p < 1$

[Saab-Yılmaz 08; Chartrand 07]

A sample of solvers

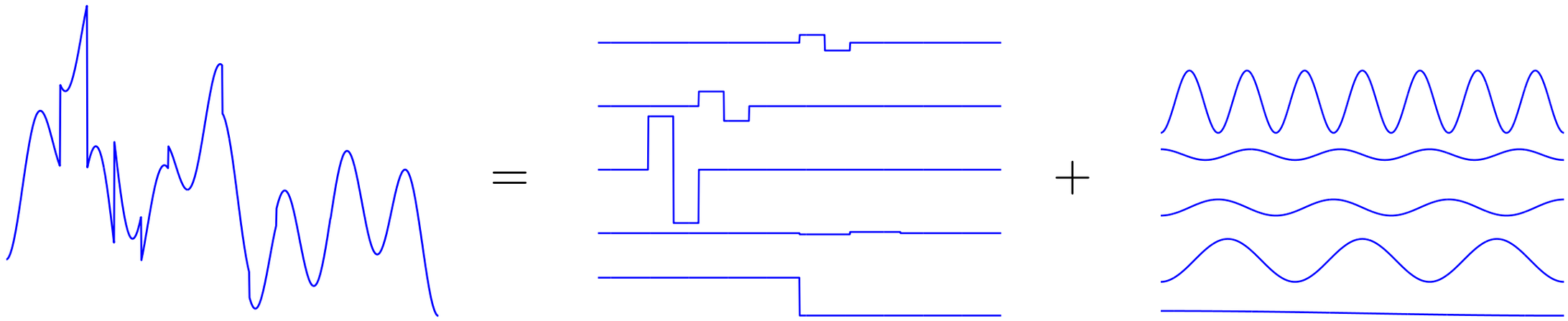
minimize $\|x\|_1$ subj to $\|Ax - b\|_2 \leq \sigma$		
ℓ_1-Magic	Candès & Romberg 05	interior, CG
SPGL1	van den Berg & Friedlander 07	root-finding
Homotopy	Osborne, Presnell & Turlach 00	active-set, all λ

minimize $\|Ax - b\|_2^2 + \lambda\|x\|_1$		
PDCO	Saunders 97, 02	interior, LSQR
L1_LS	Kim, Koh, et al 07	primal barrier, CG
BCR	Sardy, Bruce & Tseng 98	block Gauss-Seidel
GPSR	Figueiredo, Nowak & Wright 07	gradient projection
FPC	Hale, Yin & Zhang 07	operator splitting
SpaRSA	Wright, Figueiredo & Nowak 07	shrinkage

Sparse $Ax \approx b$		
OMP	Pati...93, Davis...97; Tropp & Gilbert 07	greedy
LARS	Efron, Hastie & Tibshirani 04	greedy active-set
StOMP	Donoho, Tsaig, et al 06	greedy++
ROMP	Needell & Vershynin 07	greedy+/−
CoSaMP	Needell & Tropp 08	greedy+/−

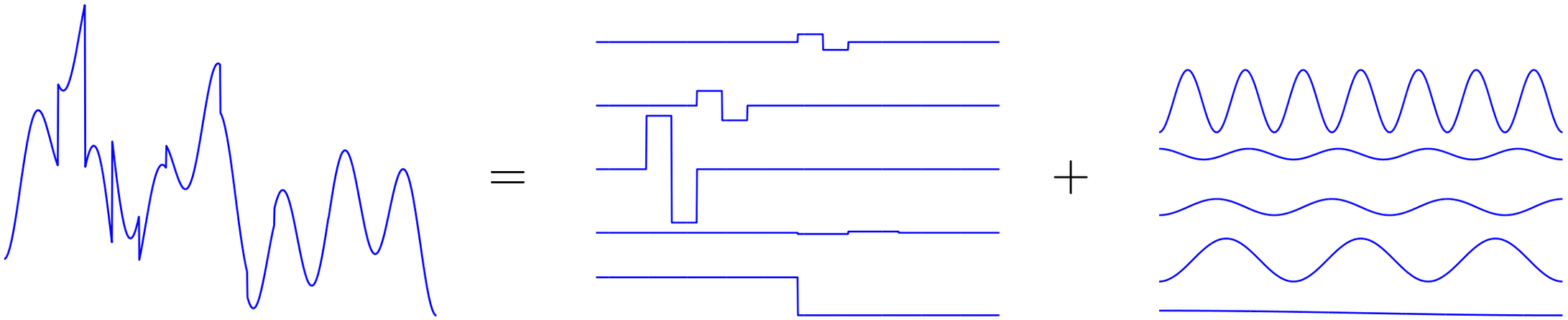
APPLICATIONS
and
FORMULATIONS

Missing observations

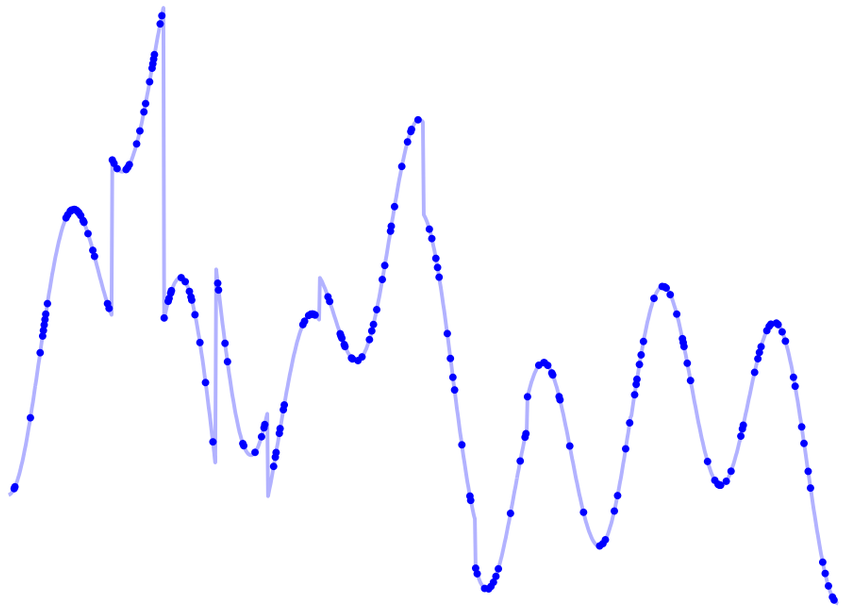


$y \approx$ Haar DCT x

Missing observations



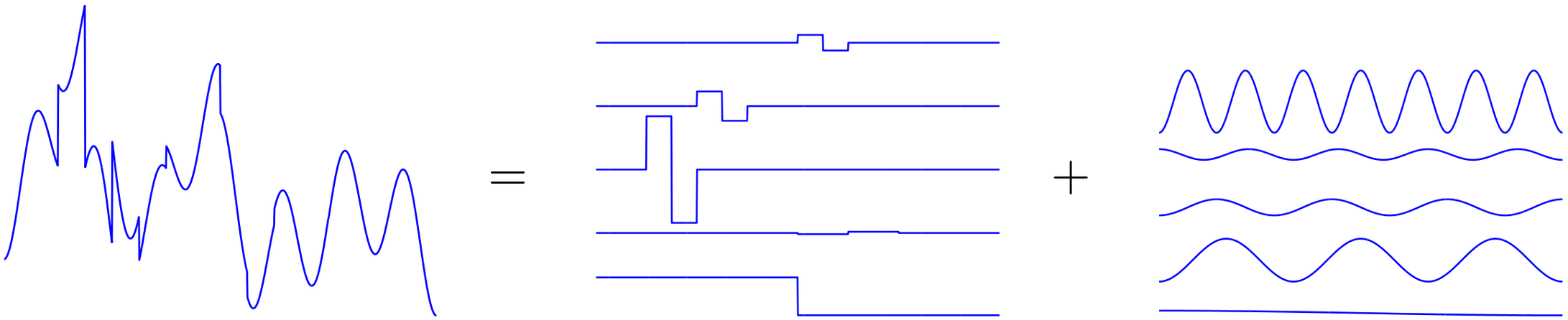
$\mathbf{R}b =$



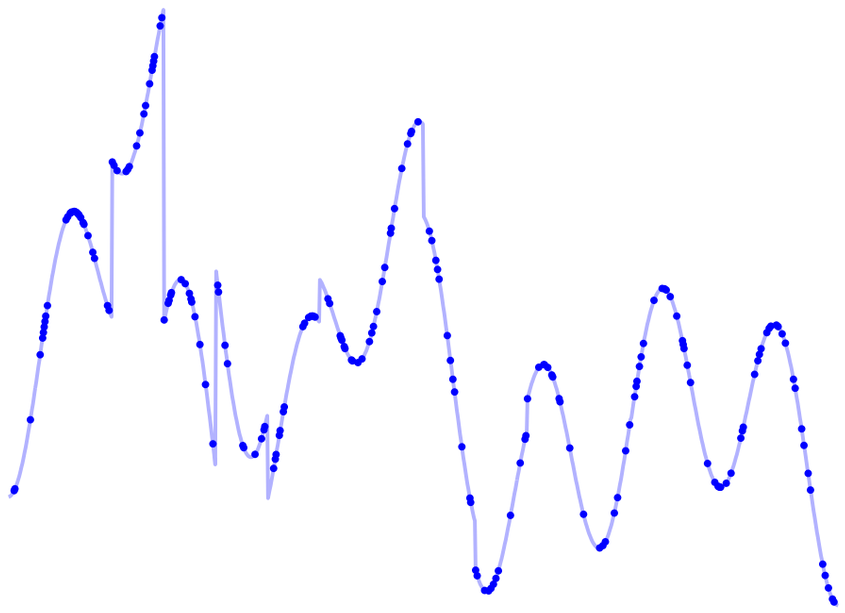
$\mathbf{R} =$ Restriction

$\mathbf{R}y \approx \mathbf{R}$ Haar DCT x

Missing observations



$\mathbf{R}b =$

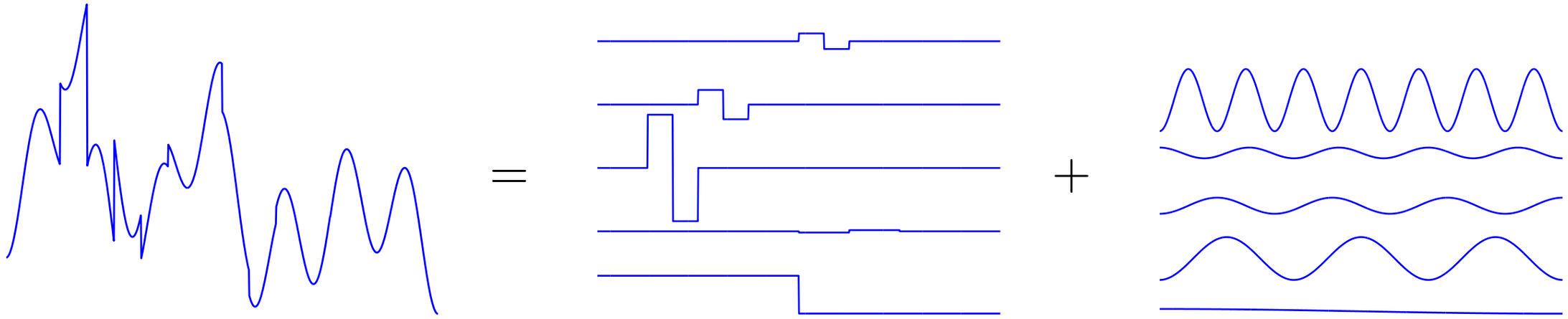


$\mathbf{R} =$ Restriction

$\mathbf{R}y \approx \mathbf{R}$ Haar DCT $x \quad \Rightarrow \quad b \approx A\,x$

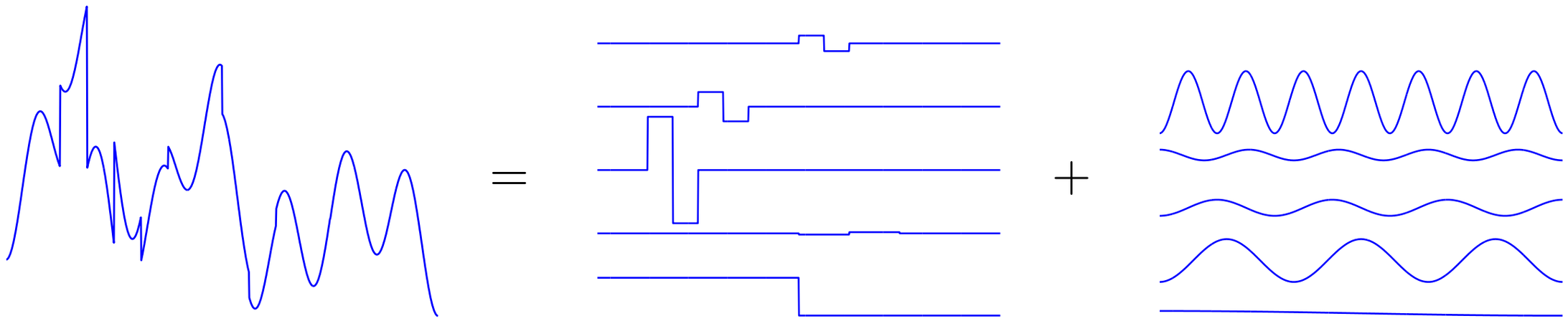
$$\text{minimize } \tfrac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_1$$

Compressed sensing

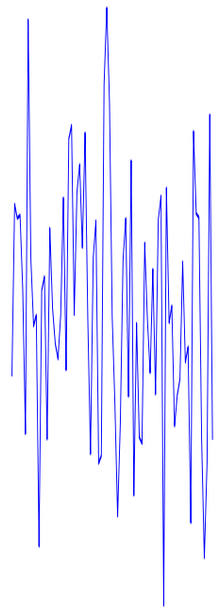


$$y \approx \boxed{\text{Haar}\ \ \text{DCT}}\, x$$

Compressed sensing



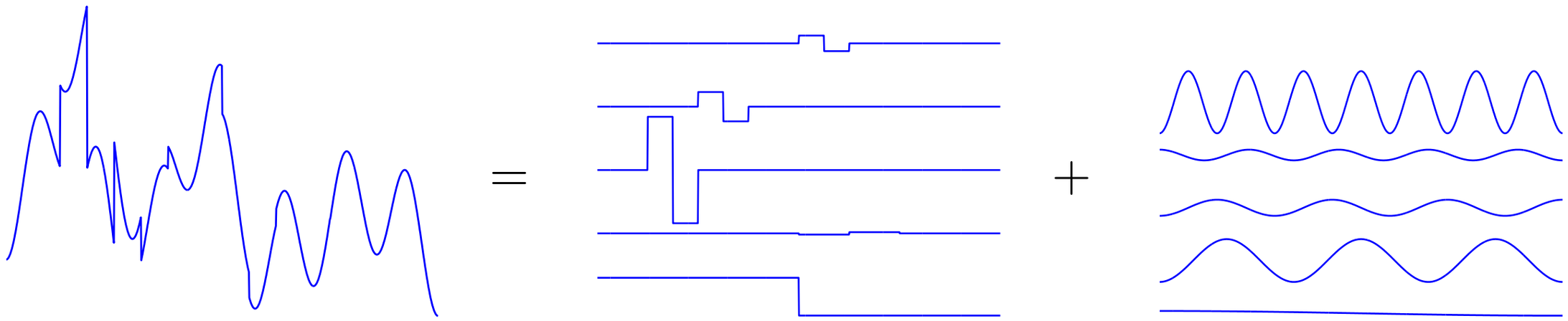
$\langle \phi_1, y \rangle, \ldots, \langle \phi_m, y \rangle$ $\quad \mathbf{\Phi} y =$



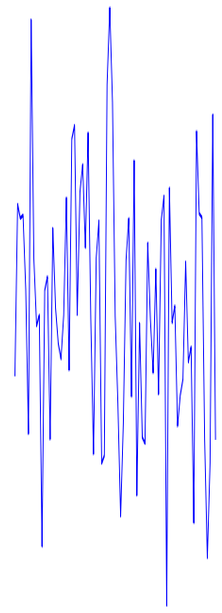
$\mathbf{\Phi} =$ Gaussian

$\mathbf{\Phi} y \approx \mathbf{\Phi}$ Haar DCT x

Compressed sensing



$$\langle \phi_1, y \rangle, \ldots, \langle \phi_m, y \rangle \qquad \mathbf{\Phi} y =$$



$$\mathbf{\Phi} = \boxed{\text{Gaussian}}$$

$$\mathbf{\Phi} y \approx \mathbf{\Phi} \boxed{\text{Haar}}\boxed{\text{DCT}} x \qquad \Rightarrow \qquad b \approx \boxed{A} x$$

$$\boxed{\text{minimize } \tfrac{1}{2}\|Ax - b\|^2 + \lambda \|x\|_1}$$

Image deblurring

Observe $b = My$, $y =$ image, $M =$ blurring operator

Image deblurring

Observe $b = My$, $y =$ image, $M =$ blurring operator

Recover significant coeff's via

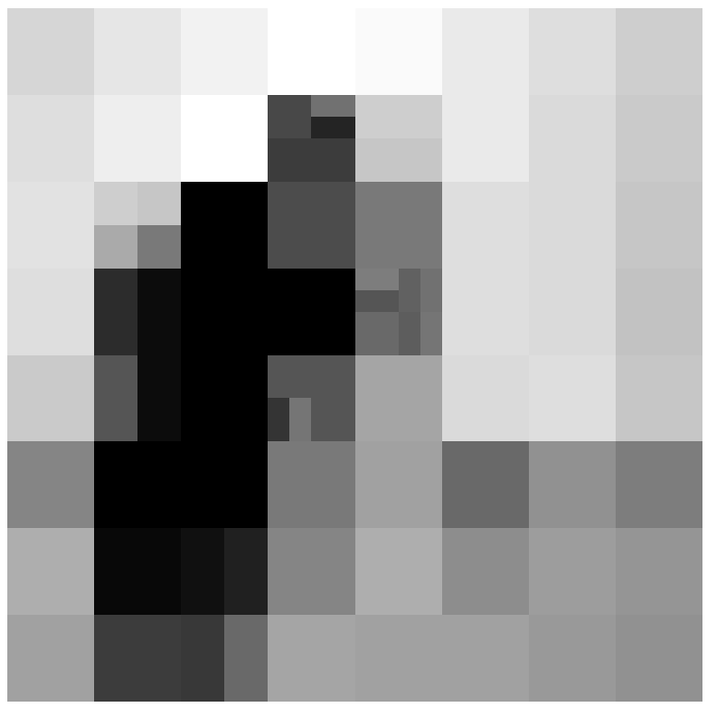
$$\text{minimize} \quad \tfrac{1}{2}\|MWx - b\|^2 + \lambda\|x\|_1$$

Image deblurring

Observe $b = My$, $y =$ image, $M =$ blurring operator

Recover significant coeff's via

$$\text{minimize} \quad \tfrac{1}{2}\|MWx - b\|^2 + \lambda\|x\|_1$$



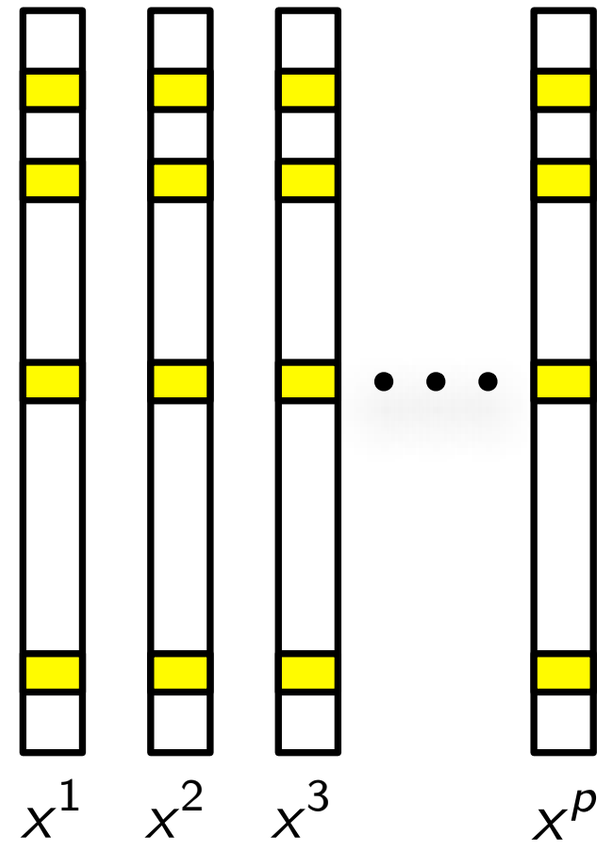
λ	2.3e+0	2.3e-2	2.3e-4
$\|Ax - b\|$	2.6e+2	3.0e+0	3.1e-2
$\text{nnz}(x)$	72	1,741	28,484

Sparco problem `blurrycam` [Figueiredo-Nowak-Wright 07]

Sum-of-norms and joint sparsity

Multiple measurements

$$Ax^1 \approx b^1, \quad Ax^2 \approx b^2, \ \ldots \ , Ax^p \approx b^p$$



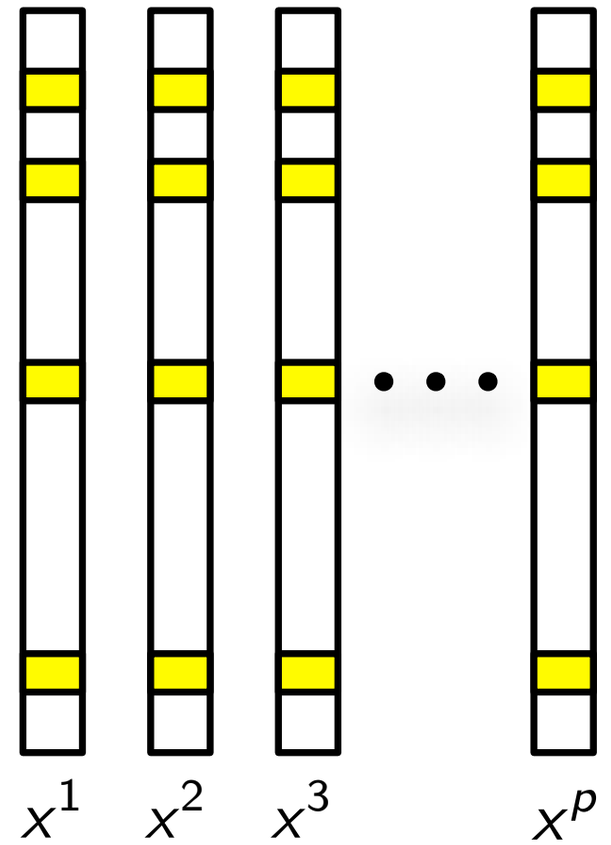
Sum-of-norms and joint sparsity

Multiple measurements

$$Ax^1 \approx b^1, \quad Ax^2 \approx b^2, \ \ldots \ , Ax^p \approx b^p$$

Minimize sum of row norms

$$\begin{array}{ll} \underset{X \in \mathbb{R}^{n \times k}}{\text{minimize}} & \|X\|_{1,2} := \sum \|X^{j\rightarrow}\|_2 \\ \text{subj to} & \|AX - B\|_F \leq \sigma \end{array}$$



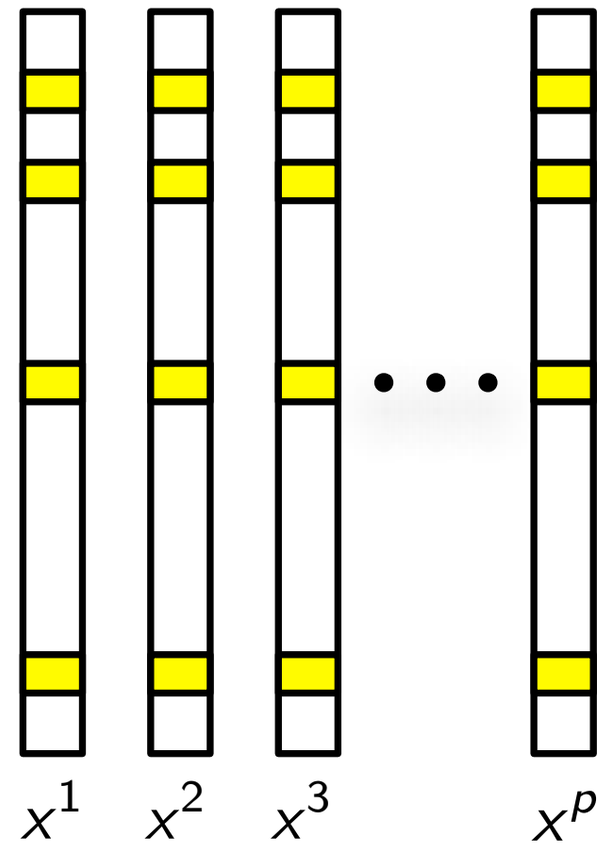
Sum-of-norms and joint sparsity

Multiple measurements

$$Ax^1 \approx b^1, \quad Ax^2 \approx b^2, \ \ldots \ , Ax^p \approx b^p$$

Minimize sum of row norms

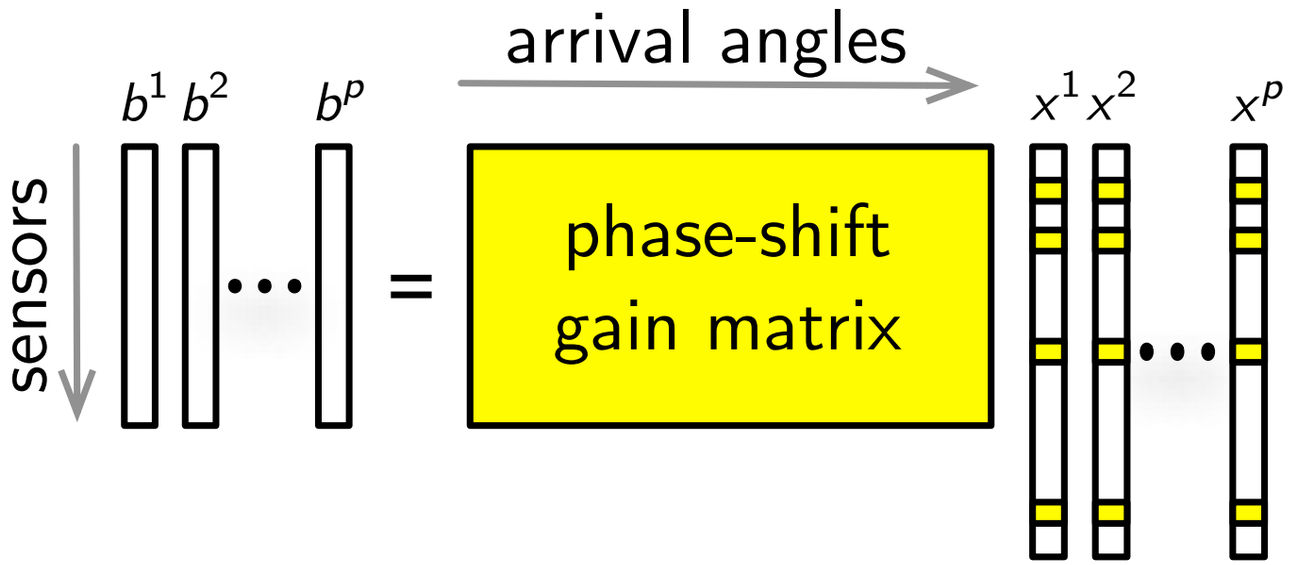
$$\begin{array}{ll} \underset{X \in \mathbb{R}^{n \times k}}{\text{minimize}} & \|X\|_{1,2} := \sum \|X^{j\rightarrow}\|_2 \\ \text{subj to} & \|AX - B\|_F \leq \sigma \end{array}$$



Projection onto $\{X \mid \|X\|_{1,2} \leq \tau\}$:

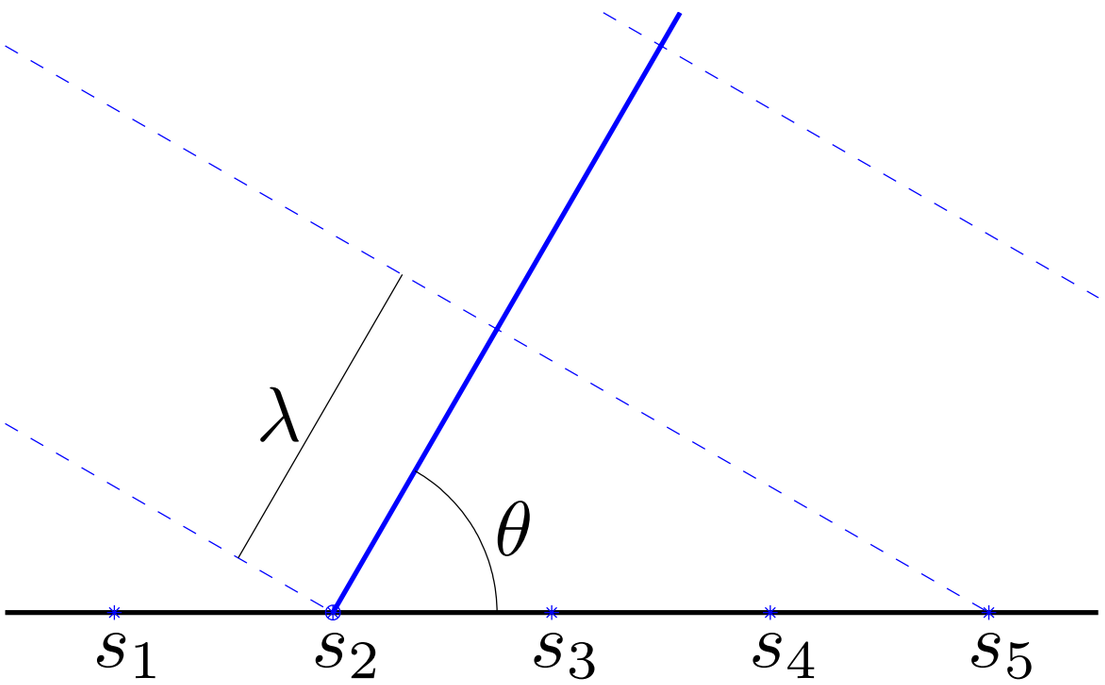
1. compute vector of row norms: $r \leftarrow (\|X^{1\rightarrow}\|_2, \ldots, \|X^{n\rightarrow}\|_2)$
2. **one-norm projection**: $\bar{r} \leftarrow \mathcal{P}_\tau[r]$
3. scale rows by $\bar{r}$: $\mathcal{P}_\tau[X] \leftarrow \text{diag}(\bar{r}./r)\ X$

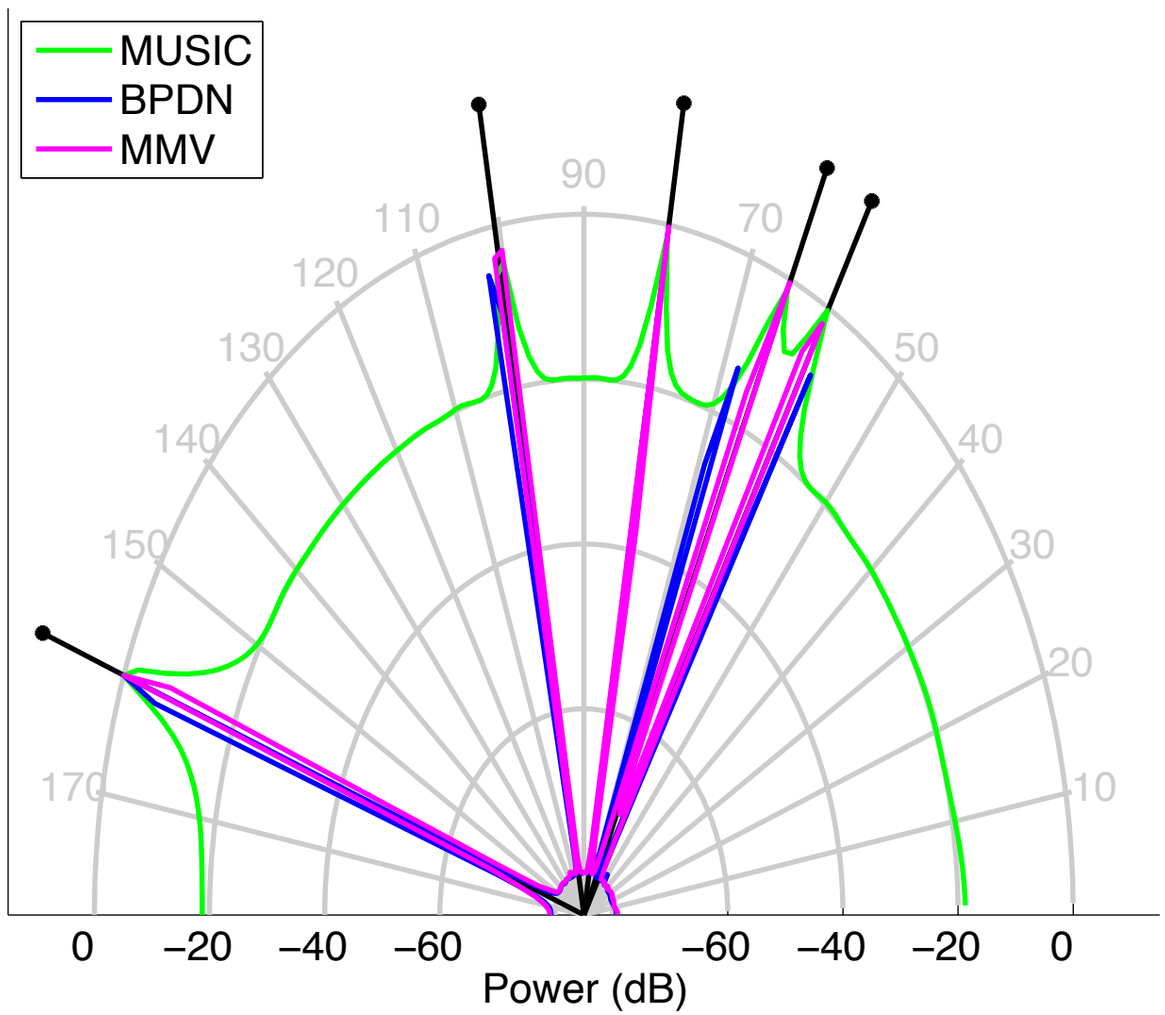
Joint sparsity: source localization



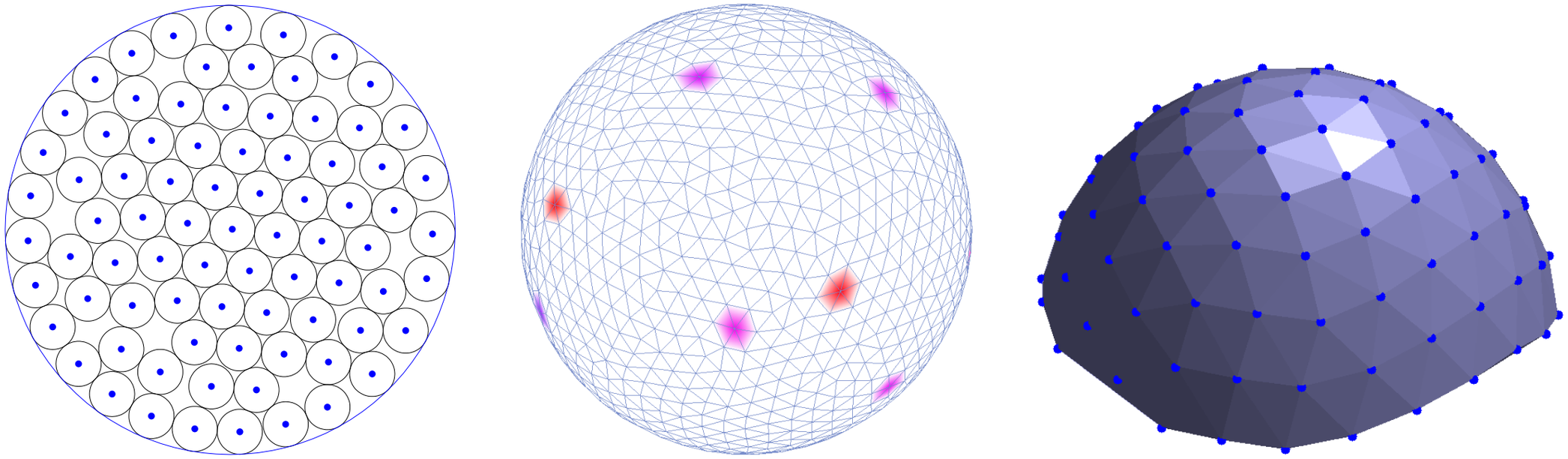
[Malioutov–Çetin–Willsky 03]

$$\begin{array}{ll} \text{minimize} & \|X\|_{1,2} \\ \text{subj to} & \|B - AX\|_F \leq \sigma \end{array}$$



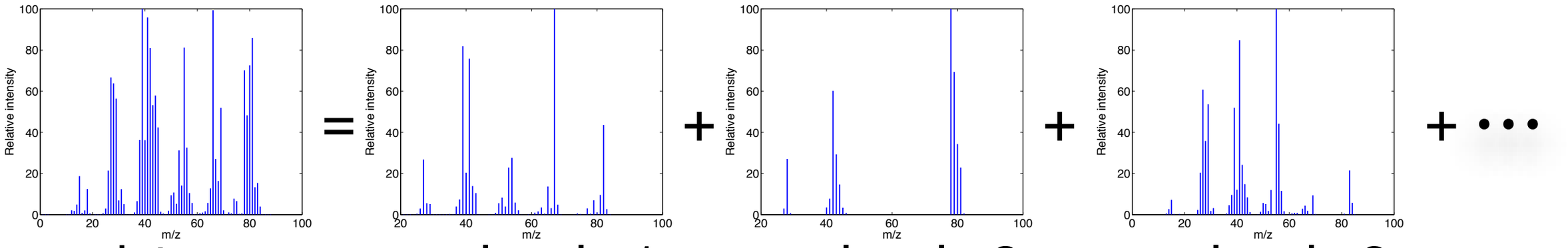


...and in 3D



[van den Berg's thesis 09]

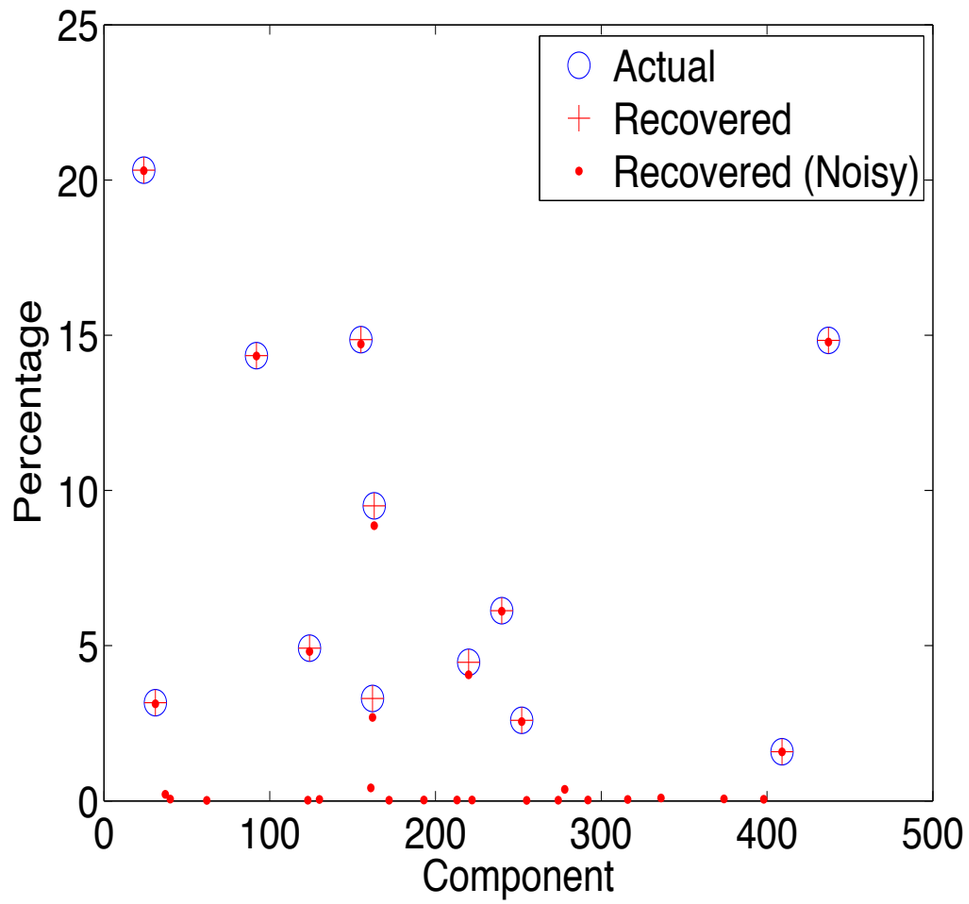
Mass spectrometry – separating mixtures



mixture molecule 1 molecule 2 molecule 3

$$A = \left| \right| \left| \right| \left| \right| \cdots \left| \right|$$

spectral fingerprints



$$\underset{x \geq 0}{\text{minimize}} \;\; \tfrac{1}{2}\|Ax - b\|^2 + \lambda \underbrace{e^T x}_{\equiv \|x\|_1}$$

[Du-Angeletti 06, vdBerg-Friedlander 10]

SPARSE OPTIMIZATION ALGORITHMS

greedy algorithms, projected gradient, iterative soft-thresholding, fixed point, proximal point, augmented Lagrangian, Bregman, split Bregman, alternating direction of method of multipliers, . . .

Canonical greedy approach

Orthogonal matching pursuit (OMP) for sparse $Ax = b$

Canonical greedy approach

Orthogonal matching pursuit (OMP) for sparse $Ax = b$

0. Initialize $r \leftarrow b, \quad S \leftarrow [\,]$

Canonical greedy approach

Orthogonal matching pursuit (OMP) for sparse $Ax = b$

0. Initialize $\quad r \leftarrow b, \quad S \leftarrow [\,]$
1. Largest correlation $\quad$ find j st $\quad |a_j^T r| = \|A^T r\|_\infty$

Canonical greedy approach

Orthogonal matching pursuit (OMP) for sparse $Ax = b$

0. Initialize $r \leftarrow b, \quad S \leftarrow [\,]$
1. Largest correlation find j st $|a_j^T r| = \|A^T r\|_\infty$
2. Get new column $S \leftarrow [\, S \quad a_j \,]$

Canonical greedy approach

Orthogonal matching pursuit (OMP) for sparse $Ax = b$

0. Initialize $\quad r \leftarrow b, \quad S \leftarrow [\,]$
1. Largest correlation $\quad \text{find} \;\; j \;\; \text{st} \;\; |a_j^T r| = \|A^T r\|_\infty$
2. Get new column $\quad S \leftarrow [\, S \quad a_j \,]$
3. Least squares $\quad \min \; \|b - Sx\|_2$

Canonical greedy approach

Orthogonal matching pursuit (OMP) for sparse $Ax = b$

0. Initialize $r \leftarrow b, \quad S \leftarrow [\,]$
1. Largest correlation find j st $|a_j^T r| = \|A^T r\|_\infty$
2. Get new column $S \leftarrow [\, S \quad a_j \,]$
3. Least squares min $\|b - Sx\|_2$
4. Update residual $r \leftarrow b - Sx$

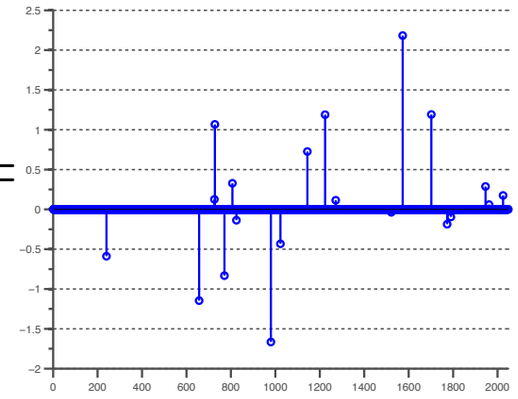
Canonical greedy approach

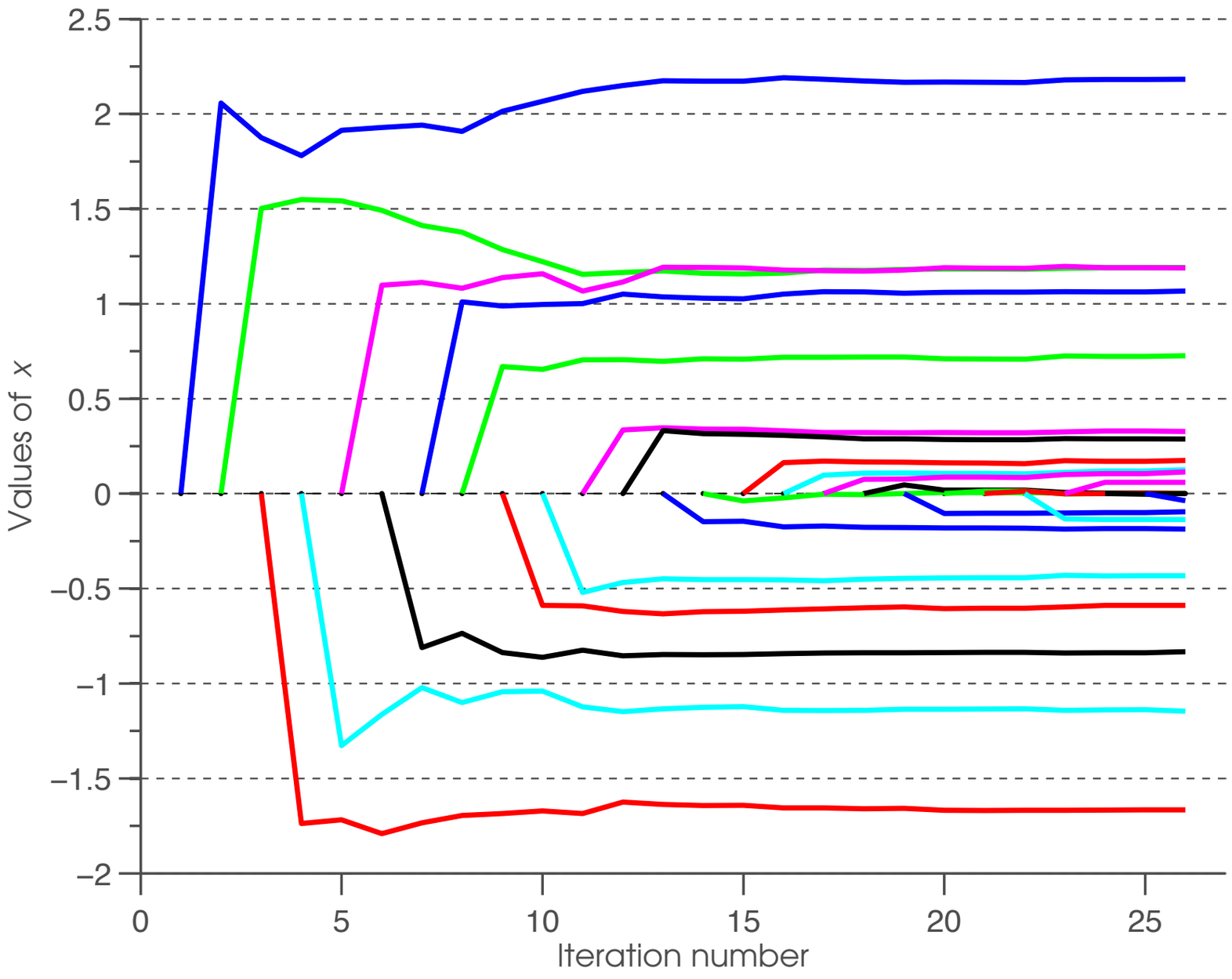
Orthogonal matching pursuit (OMP) for sparse $Ax = b$

0. Initialize $r \leftarrow b, \quad S \leftarrow [\,]$
1. Largest correlation find j st $|a_j^T r| = \|A^T r\|_\infty$
2. Get new column $S \leftarrow [\, S \quad a_j \,]$
3. Least squares min $\|b - Sx\|_2$
4. Update residual $r \leftarrow b - Sx$

- **Often** recovers sparsest solution [Tropp & Gilbert '07]
- Not optimizing a particular problem

$$Ax = b, \qquad A = \boxed{200 \times 2048}, \qquad x =$$





Forward-backward splitting

General problem:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) + P(x)$$

where

$$f \text{ is smooth}, \qquad P \text{ is sparsifying/nonsmooth}$$

Two common sparsifying P:

- $P(x) = \lambda \|x\|_1$ — minimize $f(x) + \lambda \|x\|_1$
- $P(x) = \delta_{\mathcal{C}}(x)$ with $\mathcal{C} = \{x \mid \|x\|_1 \leq \tau\}$ — minimize $f(x)$ st $\|x\|_1 \leq \tau$

Many related methods:

projected gradient, iterative soft-thresholding, fixed point, proximal point, augmented Lagrangian, Bregman, split Bregman, alternating direction method of multipliers, etc

Proximity operators

Proximity operator of P:

$$\textbf{prox}_P(x) \quad \text{solves} \quad \underset{z \in \mathbb{R}^n}{\text{minimize}} \ \tfrac{1}{2}\|x - z\|^2 + P(z)$$

Generalizes the **projection** operator of $\mathcal{C} \subseteq \mathbb{R}^n$:

$$\textbf{proj}_{\mathcal{C}}(x) \quad \text{solves} \quad \underset{z \in \mathbb{R}^n}{\text{minimize}} \ \tfrac{1}{2}\|x - z\|^2 + \delta_{\mathcal{C}}(z)$$

where $\delta_{\mathcal{C}}(x)$ is the indicator function for $\mathcal{C}$:

$$\delta_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{C} \\ \infty & \text{otherwise} \end{cases}$$

Soft thresholding is proximity operator of the one-norm:

$$P(z) = \|z\|_1 \quad \rightarrow \quad \textbf{prox}_P(x) := \mathcal{S}_\lambda(x) := \text{sgn}(x) \cdot \max\{|x| - \lambda, 0\}$$

Forward-backward splitting algorithm

$$\boxed{\text{minimize } f(x) + P(x)}$$

Optimality: x^* is a solution if and only if

$$x^* = \text{prox}_{\alpha P}(x^* - \alpha \nabla f(x^*)) \quad \text{for all } \alpha > 0$$

Algorithm:

$$x^{k+1} = \underbrace{\text{prox}_{\alpha^k P}}_{\substack{\text{backward} \\ \text{step}}} \Big[\underbrace{x^k - \alpha^k \nabla f(x^k)}_{\substack{\text{forward} \\ \text{step}}} \Big]$$

where

$$\alpha^k \in (\mu, 2/L), \quad \mu > 0, \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

Convergence:

Every sequence x^k generated by this algorithm converges to a solution

Iterative soft-thresholding (IST)

$$\boxed{\text{minimize } f(x) + \lambda\|x\|_1}$$

For $f(x) = \frac{1}{2}\|r\|^2$, $r \equiv Ax - b$,

$$\begin{aligned} x^{k+1} &= \text{prox}_{\lambda\|\cdot\|_1}(x^k - \alpha^k A^T r_k) \\ &= \mathcal{S}_\lambda(x^k - \alpha^k A^T r_k) \end{aligned}$$

where

$$\alpha^k \in (0, 2/\|A\|^2) \quad \text{and} \quad \mathcal{S}_\lambda(x) = \text{sgn}(x) \cdot \max\{|x| - \lambda, 0\}$$

[aka, iterative shrinkage, iterative Landweber]

Derived from various viewpoints:

- expectation-maximization [Figueiredo & Nowak 03]
- surrogate functionals [Daubechies et al 04]
- forward-backward splitting: min $f(x) + P(x)$ [Combettes & Wajs 05]
- fixed-point w/continuation: min $f(x) + \lambda\|x\|_1$ [Hale et al 07]

Research & Development

Research and development

- Hybrid stochastic-deterministic methods

  Mark Schmidt, PDF

- Weighted one-norm minimization

  Hassan Mansour, PDF

- Newton methods for sparse optimization

  Ives Macedo, incoming PhD

- Spot: A linear-operator toolbox for Matlab

  Ewout van den Berg (2009 PhD), et al.

Email

- `mpf@cs.ubc.ca`

Web

- `http://www.cs.ubc.ca/~mpf`