Review

Optimization-Based Data Analysis

http://www.cims.nyu.edu/~cfgranda/pages/OBDA_spring16

Carlos Fernandez-Granda

5/9/2016

How to deal with a data-analysis problem

1. Define the problem
2. Establish assumptions on signal structure
3. Design an efficient algorithm
4. Understand under what conditions the problem is well posed
5. Derive theoretical guarantees

Data-analysis problems

Signal structure

Methods

- General techniques
- Denoising
- Signal recovery
- Signal separation
- Regression
- Compression / dimensionality reduction
- Clustering

When is the problem well posed?

Theoretical analysis

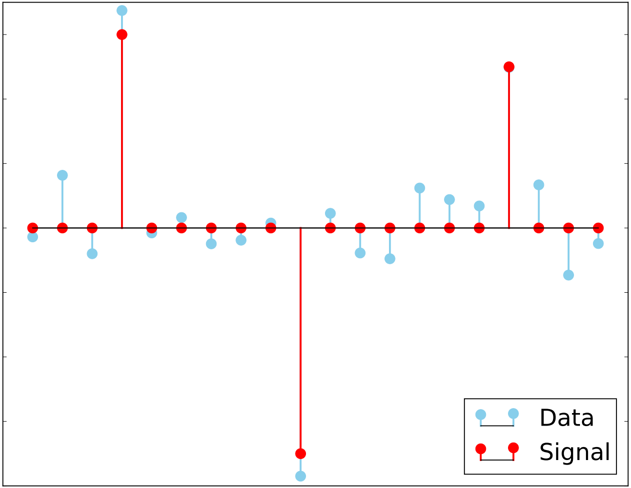
Denoising

Aim: Extracting information (signal) from data in the presence of uninformative perturbations (noise)

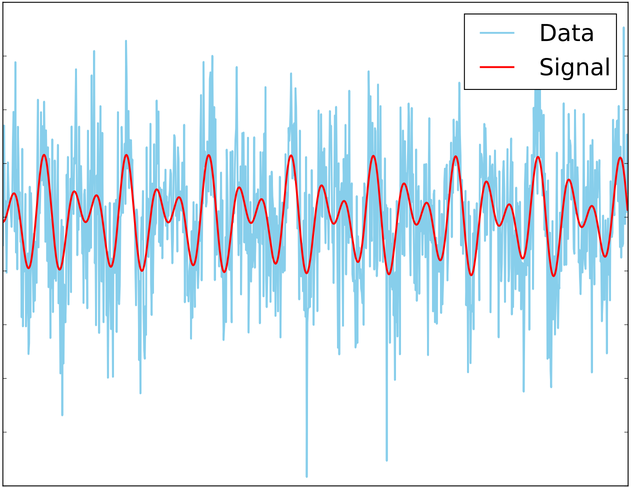
Additive noise model

$$\text{data} = \text{signal} + \text{noise}$$
$$y = x + z$$

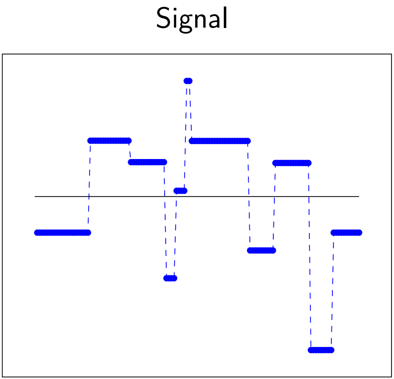
Denoising

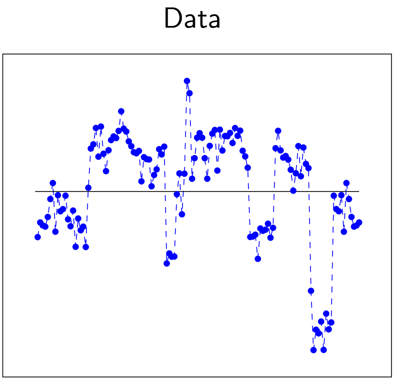


Denoising



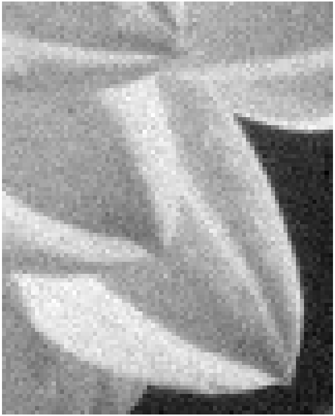
Denoising



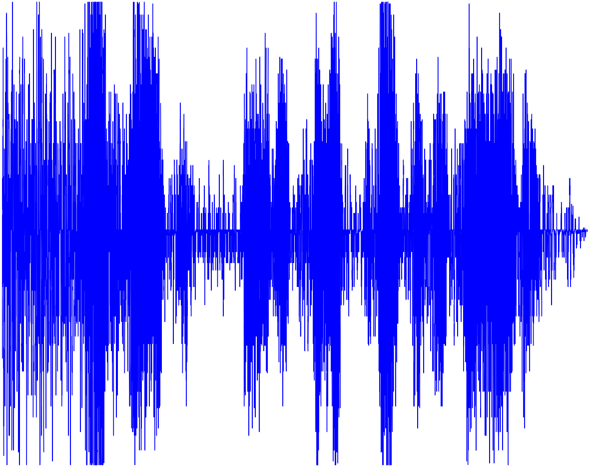


Denoising

Denoising



Denoising



Signal recovery

- Compressed sensing
- Deconvolution / super-resolution
- Matrix completion

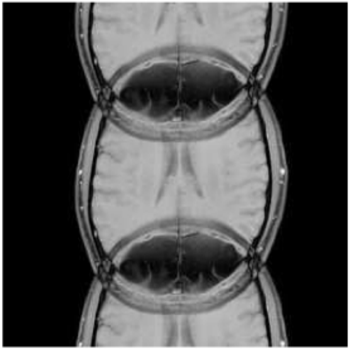
General model

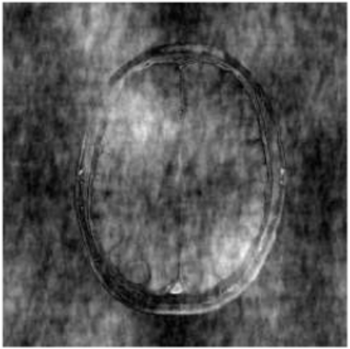
Aim: Estimate signal x from measurements y

$$y = A\,x$$

Linear underdetermined system where dimension $(y) <$ dimension (x)

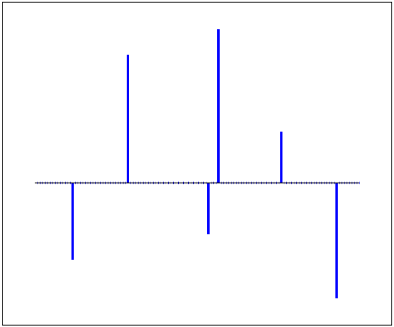
Compressed sensing



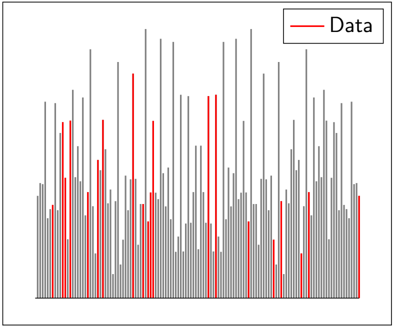


Compressed sensing

Signal

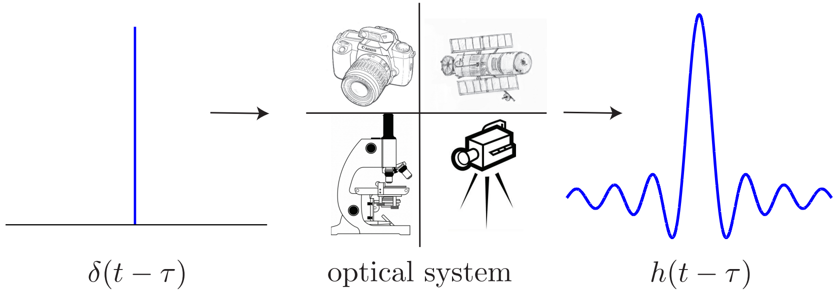


Spectrum



Super-resolution

The resolving power of lenses, however perfect, is limited (**Lord Rayleigh**)



Diffraction imposes a fundamental limit on the resolution of optical systems

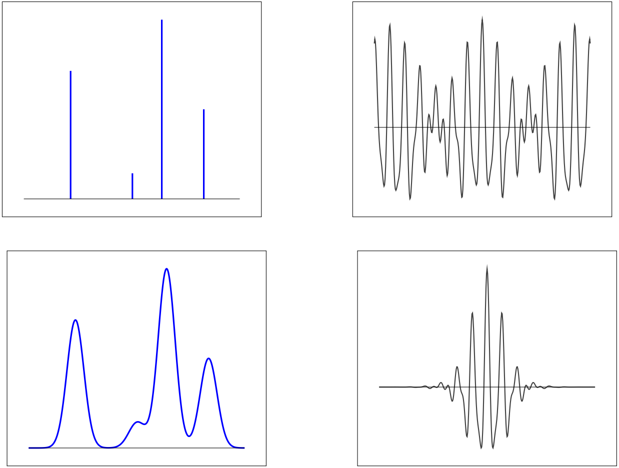
Super-resolution

Spatial Super-resolution

Spectrum

Signal

Data

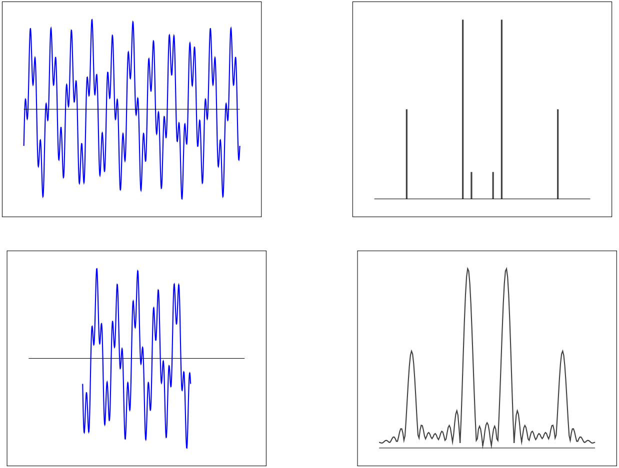


Spectral Super-resolution

Spectrum

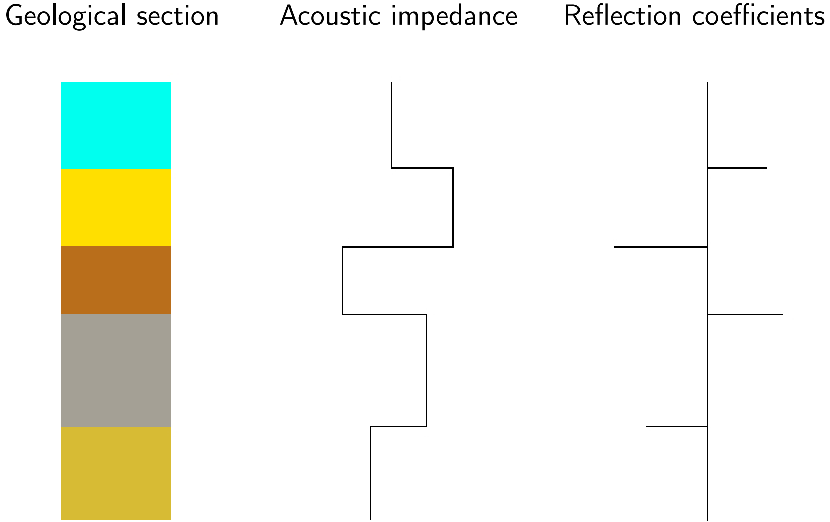
Signal

Data

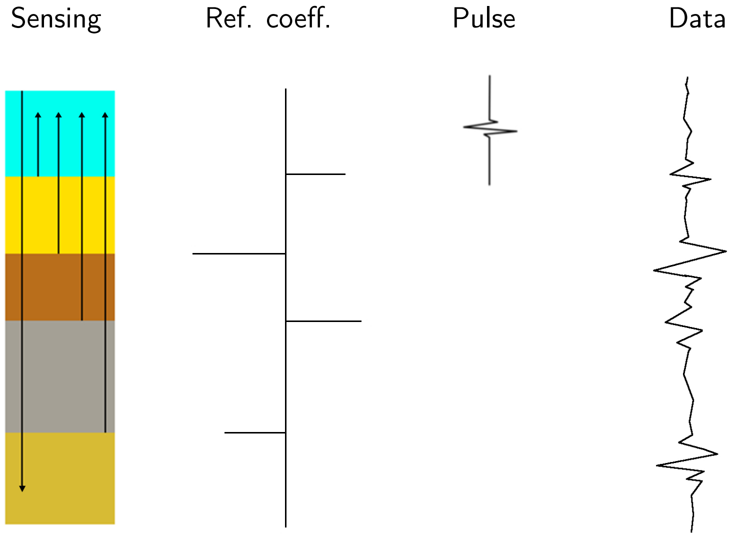


Seismology

Reflection seismology

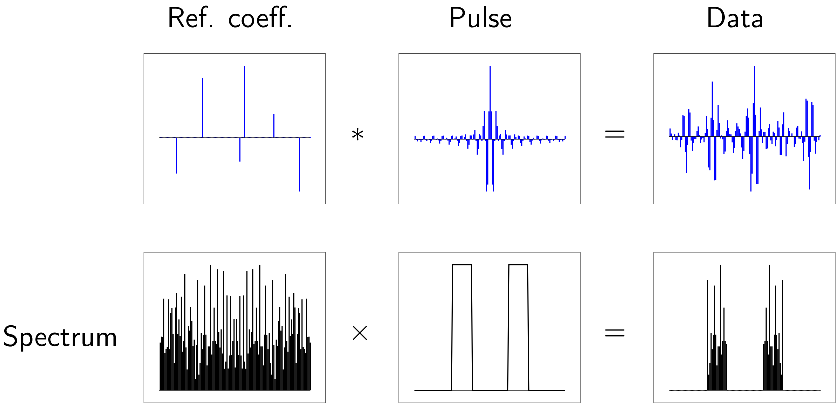


Deconvolution



Data ≈ convolution of pulse and reflection coefficients

Deconvolution



Matrix completion

Matrix completion

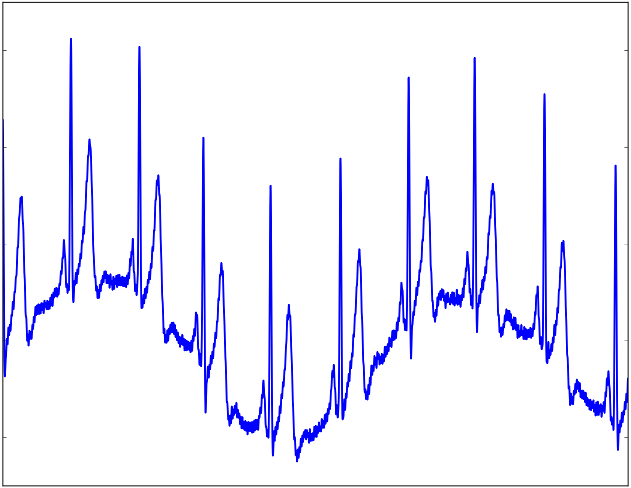
$$\begin{array}{cl} \begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} & \\ \begin{pmatrix} 1 & ? & 5 & 4 \\ ? & 1 & 4 & 5 \\ 4 & 5 & 2 & ? \\ 5 & 4 & 2 & 1 \\ 4 & 5 & 1 & 2 \\ 1 & 2 & ? & 5 \end{pmatrix} & \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array} \end{array}$$

Signal separation

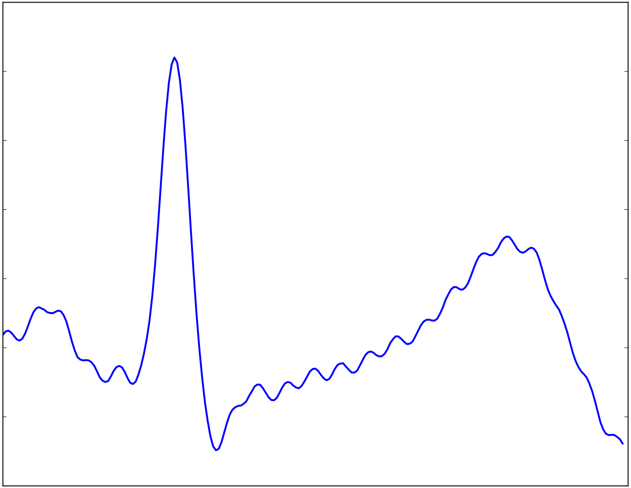
Aim: Decompose the data into two (or more) signals

$$y = x_1 + x_2$$

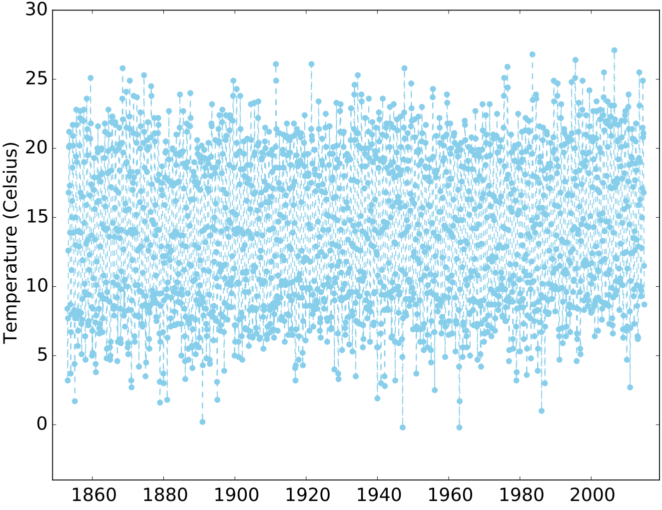
Electrocardiogram



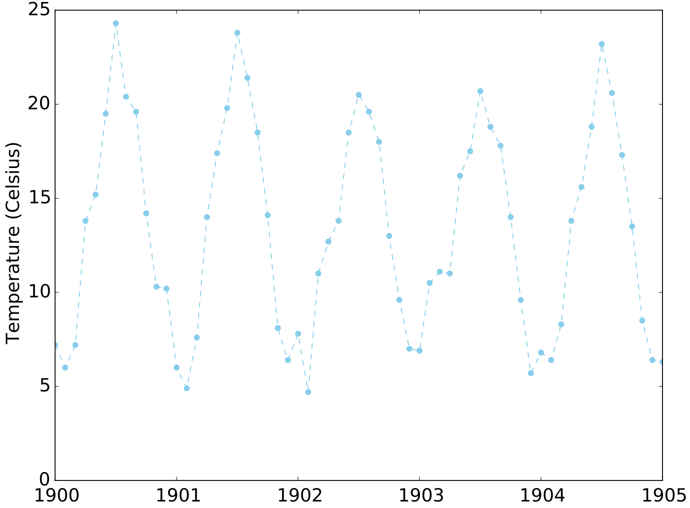
Electrocardiogram



Temperature data

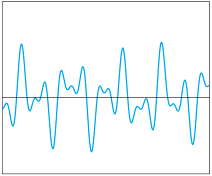


Temperature data

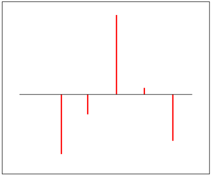


Demixing of sines and spikes

Sines



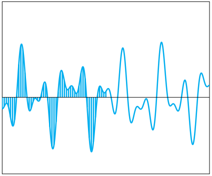
Spectrum



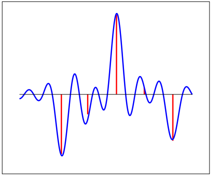
x

Demixing of sines and spikes

Sines



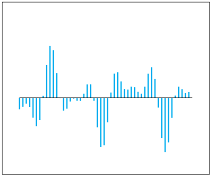
Spectrum



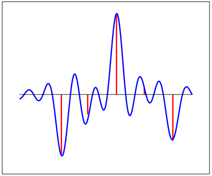
$\mathcal{F}_c\, x$

Demixing of sines and spikes

Sines



Spectrum

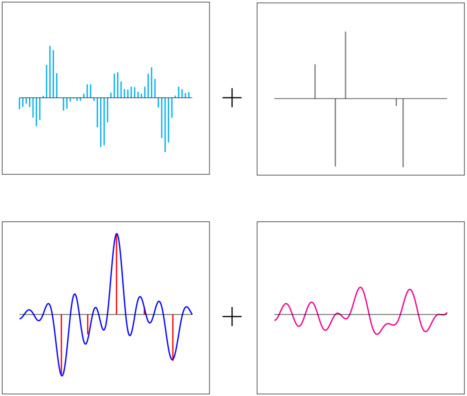


$\mathcal{F}_c\, x$

Demixing of sines and spikes

Sines | Spikes

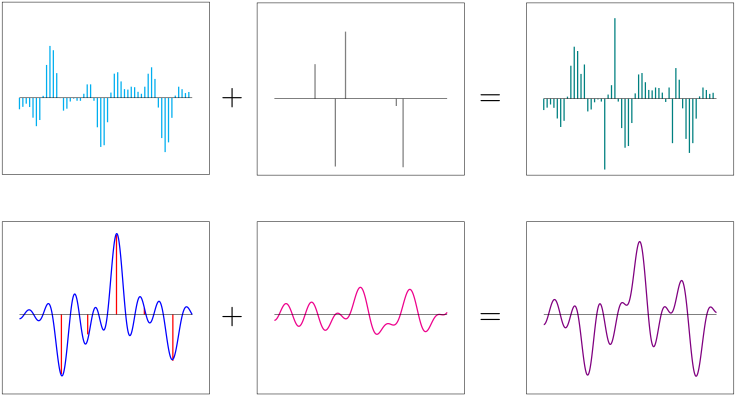
Spectrum



$$\mathcal{F}_c\, x \quad + \quad s$$

Demixing of sines and spikes

Sines | Spikes | Data



Spectrum

$$\mathcal{F}_c\, x \quad + \quad s \quad = \quad y$$

Collaborative filtering with outliers

$$A := \begin{array}{cl} \begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} & \\ \left(\begin{array}{cccc} 5 & 1 & 5 & 5 \\ 1 & 1 & 5 & 5 \\ 5 & 5 & 1 & 1 \\ 5 & 5 & 1 & 1 \\ 5 & 5 & 1 & 1 \\ 1 & 1 & 5 & 1 \end{array}\right) & \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array} \end{array}$$

Background subtraction

L

$+$

S

$=$

M

Regression

Aim: Predict the value of a response $y \in \mathbb{R}$ from p predictors $X_1, X_2, \ldots, X_p \in \mathbb{R}$

Methodology:

1. Fit a model with using n training examples $y_1, y_2, \ldots, y_n$

$$y_i \approx f\left(X_{i1}, X_{i2}, \ldots, X_{ip}\right) \quad 1 \leq i \leq n$$

2. Use learned model f to predict from new data

Sparse regression

Assumption: Response only depends on a subset $\mathcal{S}$ of $s \ll p$ predictors

Model-selection problem: Determine what predictors are relevant

Classification

Aim: Predict the value of a binary response $y \in \{0, 1\}$ from p predictors X_1, X_2, ..., $X_p \in \mathbb{R}$

Methodology:

1. Fit a model with using n training examples y_1, y_2, ..., y_n

$$y_i \approx f\left(X_{i1}, X_{i2}, \ldots, X_{ip}\right) \quad 1 \leq i \leq n$$

2. Use learned model f to predict from new data

Arrhythmia prediction

Predict whether patient has arrhythmia from $n = 271$ examples and $p = 182$ predictors

- Age, sex, height, weight
- Features obtained from electrocardiogram recordings

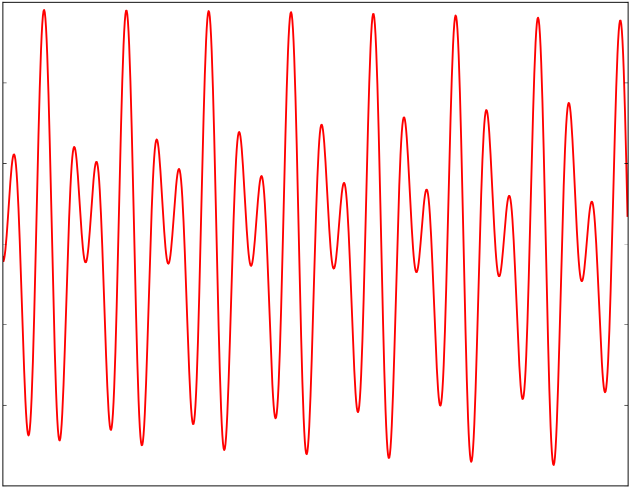
Compression

Aim: Map a signal $x \in \mathbb{R}^n$ to a lower-dimensional space

$$y \approx f(x)$$

such that we can recover x from y with minimal loss of information

Compression



Compression

Dimensionality reduction

Projection of data onto lower-dimensional space

- Decreases computational cost of processing the data
- Allows to visualize (2D, 3D)

Difference with compression: Not necessarily reversible

Dimensionality reduction

Seeds from three different varieties of wheat: Kama, Rosa and Canadian

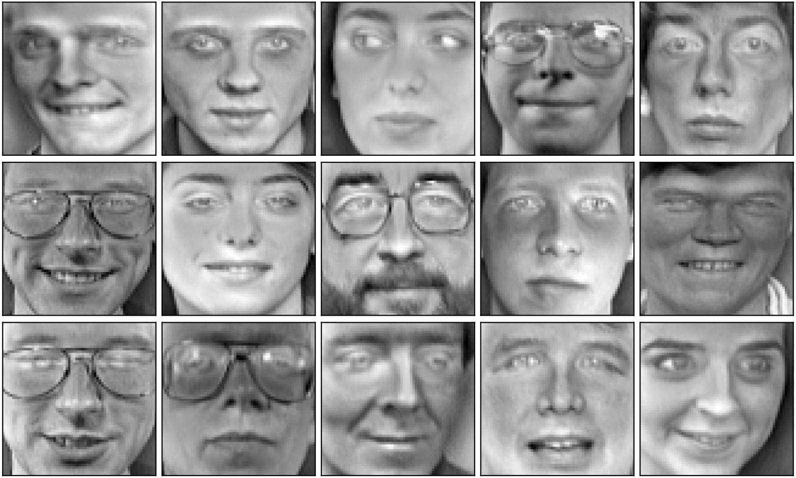
Dimensions:

- Area
- Perimeter
- Compactness
- Length of kernel
- Width of kernel
- Asymmetry coefficient
- Length of kernel groove

Clustering

Aim: Separate signals $x_1, \ldots, x_n \in \mathbb{R}^d$ into different classes

Clustering



Collaborative filtering

$$A := \begin{array}{c@{}l} \begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} & \\ \left(\begin{array}{cccc} 1 & 1 & 5 & 4 \\ 2 & 1 & 4 & 5 \\ 4 & 5 & 2 & 1 \\ 5 & 4 & 2 & 1 \\ 4 & 5 & 1 & 2 \\ 1 & 2 & 5 & 5 \end{array}\right) & \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array} \end{array}$$

Topic modeling

$$A := \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array}\begin{array}{c} \begin{array}{ccccccccc} \text{singer} & \text{GDP} & \text{senate} & \text{election} & \text{vote} & \text{stock} & \text{bass} & \text{market} & \text{band} \end{array} \\ \begin{pmatrix} 6 & 1 & 1 & 0 & 0 & 1 & 9 & 0 & 8 \\ 1 & 0 & 9 & 5 & 8 & 1 & 0 & 1 & 0 \\ 8 & 1 & 0 & 1 & 0 & 0 & 9 & 1 & 7 \\ 0 & 7 & 1 & 0 & 0 & 9 & 1 & 7 & 0 \\ 0 & 5 & 6 & 7 & 5 & 6 & 0 & 7 & 2 \\ 1 & 0 & 8 & 5 & 9 & 2 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{array}{c} \text{Articles} \\ \text{a} \\ \text{b} \\ \text{c} \\ \text{d} \\ \text{e} \\ \text{f} \end{array}$$

Models

- Sparse models
- Group sparse models
- Low-rank models

Sparsity

$$x = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 3 \\ 0 \end{bmatrix}$$

Group sparsity

Entries are partitioned into m groups $\mathcal{G}_1$, $\mathcal{G}_2$, ..., $\mathcal{G}_m$

$$x = \begin{bmatrix} x_{\mathcal{G}_1} \\ x_{\mathcal{G}_2} \\ \cdots \\ x_{\mathcal{G}_m} \end{bmatrix}$$

Assumption: Most groups are zero

Sparse models

Let D be a dictionary of atoms

1. Synthesis sparse model

$$x = Dc \quad \text{where } c \text{ is sparse}$$

2. Analysis sparse model:

$$D^T x \text{ is sparse}$$

Low-rank model

Signal is structured as a matrix that presents significant correlations

Collaborative filtering

$$
A := \begin{array}{c}
\begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} \\
\left(\begin{array}{cccc}
1 & 1 & 5 & 4 \\
2 & 1 & 4 & 5 \\
4 & 5 & 2 & 1 \\
5 & 4 & 2 & 1 \\
4 & 5 & 1 & 2 \\
1 & 2 & 5 & 5
\end{array} \right)
\end{array}
\begin{array}{l}
\\
\text{The Dark Knight} \\
\text{Spiderman 3} \\
\text{Love Actually} \\
\text{Bridget Jones's Diary} \\
\text{Pretty Woman} \\
\text{Superman 2}
\end{array}
$$

SVD

$$A - \bar{A} = U \Sigma V^T = U \begin{bmatrix} 7.79 & 0 & 0 & 0 \\ 0 & 1.62 & 0 & 0 \\ 0 & 0 & 1.55 & 0 \\ 0 & 0 & 0 & 0.62 \end{bmatrix} V^T$$

Topic modeling

$$A := \begin{array}{c} \begin{array}{ccccccccc} \text{singer} & \text{GDP} & \text{senate} & \text{election} & \text{vote} & \text{stock} & \text{bass} & \text{market} & \text{band} \end{array} \\ \begin{pmatrix} 6 & 1 & 1 & 0 & 0 & 1 & 9 & 0 & 8 \\ 1 & 0 & 9 & 5 & 8 & 1 & 0 & 1 & 0 \\ 8 & 1 & 0 & 1 & 0 & 0 & 9 & 1 & 7 \\ 0 & 7 & 1 & 0 & 0 & 9 & 1 & 7 & 0 \\ 0 & 5 & 6 & 7 & 5 & 6 & 0 & 7 & 2 \\ 1 & 0 & 8 & 5 & 9 & 2 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{array}{c} \text{Articles} \\ \text{a} \\ \text{b} \\ \text{c} \\ \text{d} \\ \text{e} \\ \text{f} \end{array}$$

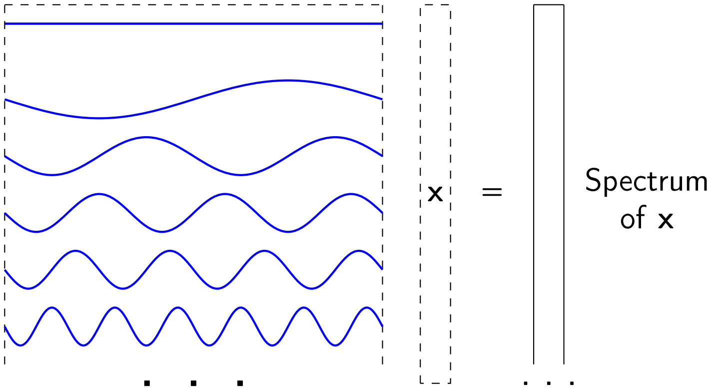
SVD

$$A = U \Sigma V^T = U \begin{bmatrix} 23.64 & 0 & 0 & 0 & & \\ 0 & 18.82 & 0 & 0 & 0 & 0 \\ 0 & 0 & 14.23 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.63 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.03 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.36 \end{bmatrix} V^T$$

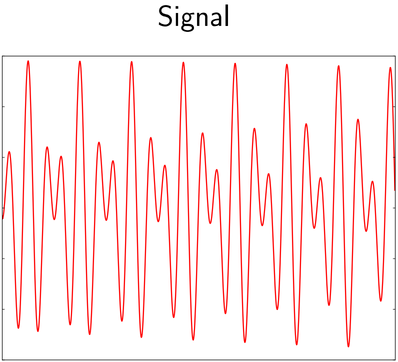
Designing signal representations

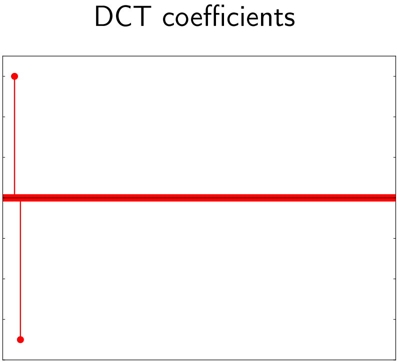
- Frequency representation
- Short-time Fourier transform
- Wavelets
- Finite differences

Frequency representation

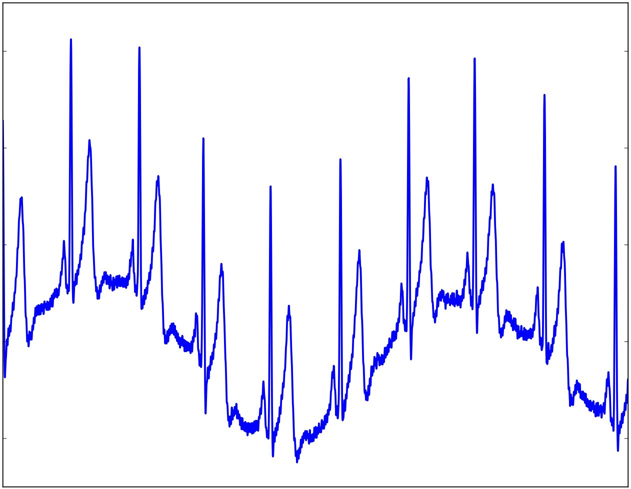


Discrete cosine transform

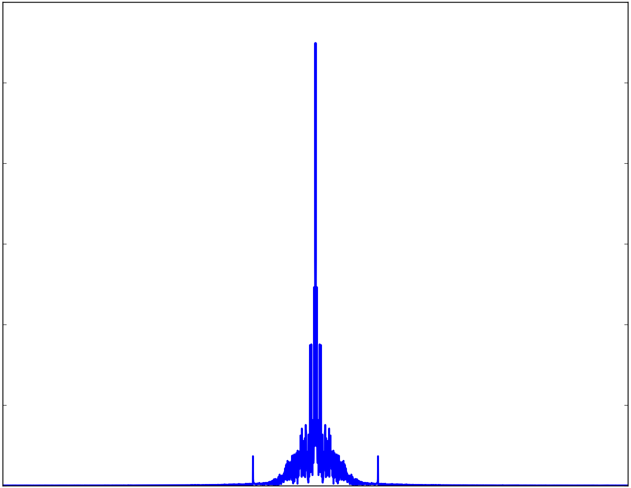




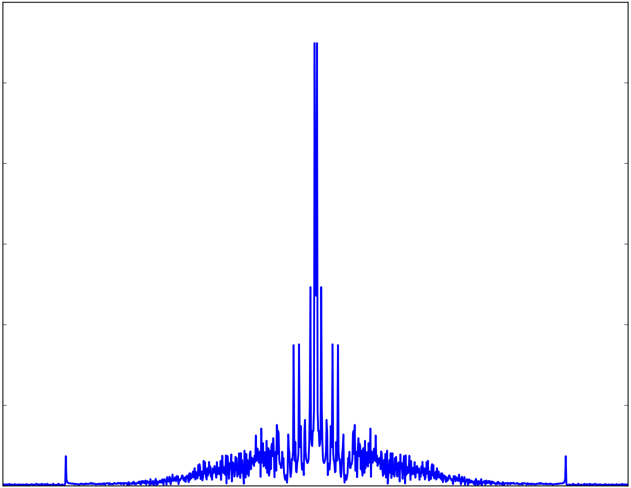
Electrocardiogram



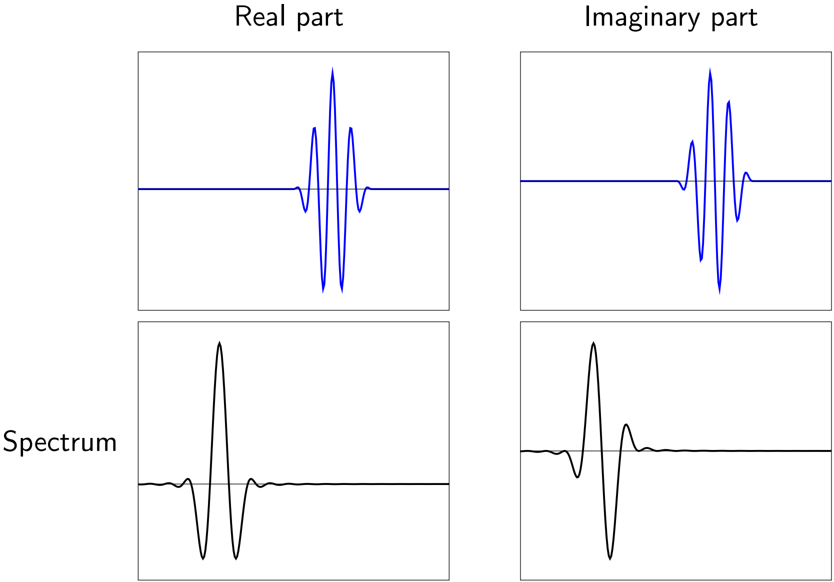
Electrocardiogram (spectrum)



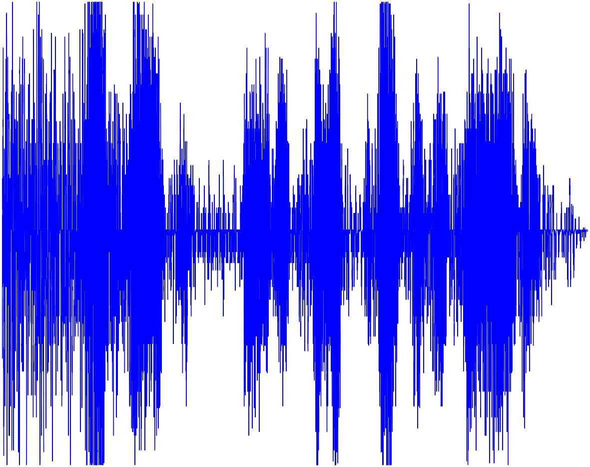
Electrocardiogram (spectrum)



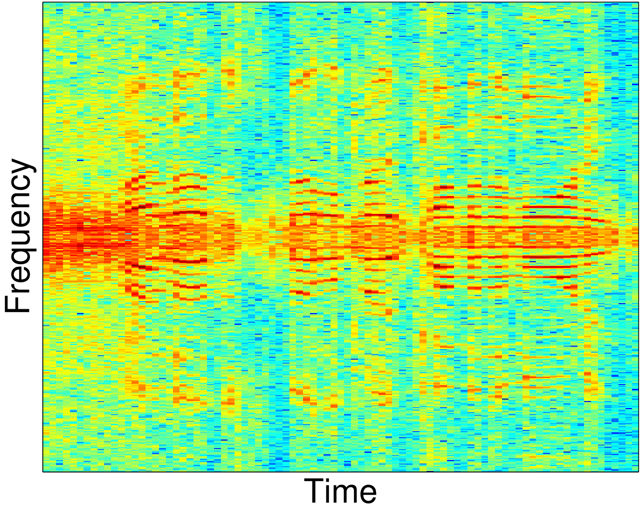
Short-time Fourier transform



Speech signal

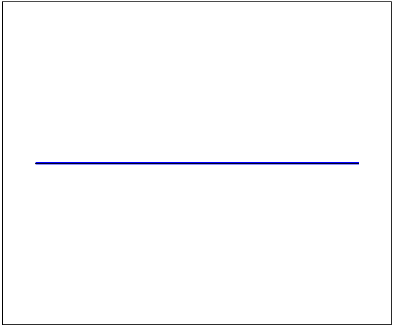


Spectrogram (log magnitude of STFT coefficients)

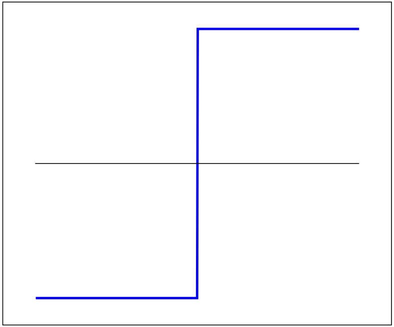


Wavelets

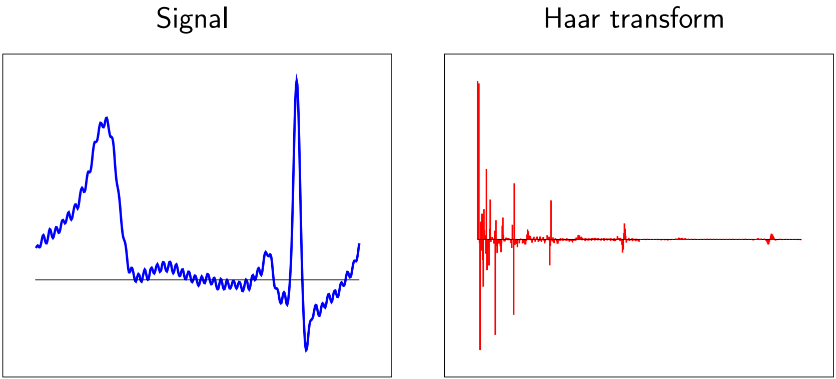
Scaling function



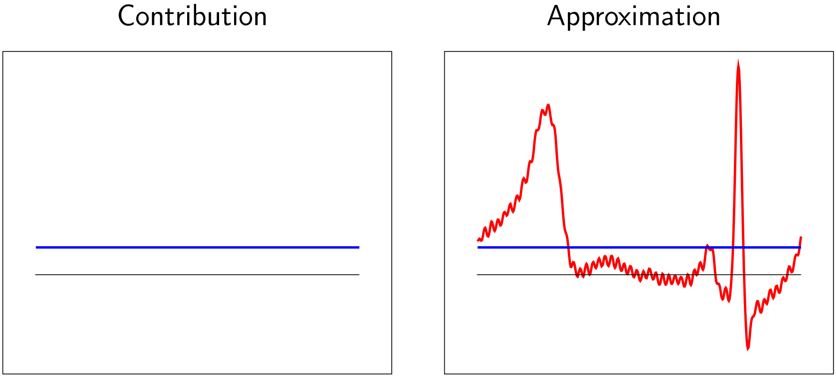
Mother wavelet



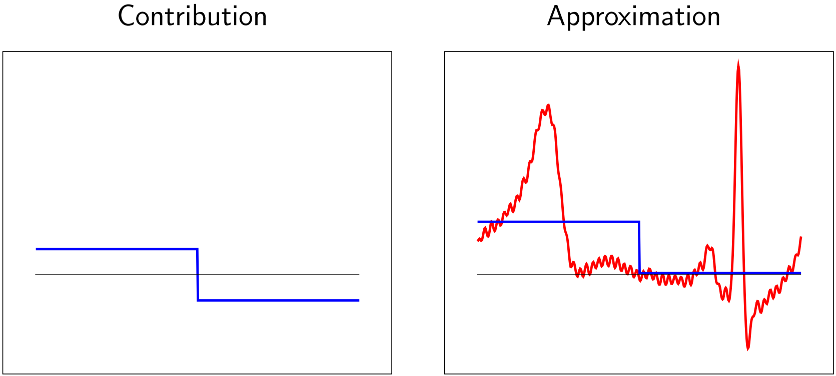
Electrocardiogram



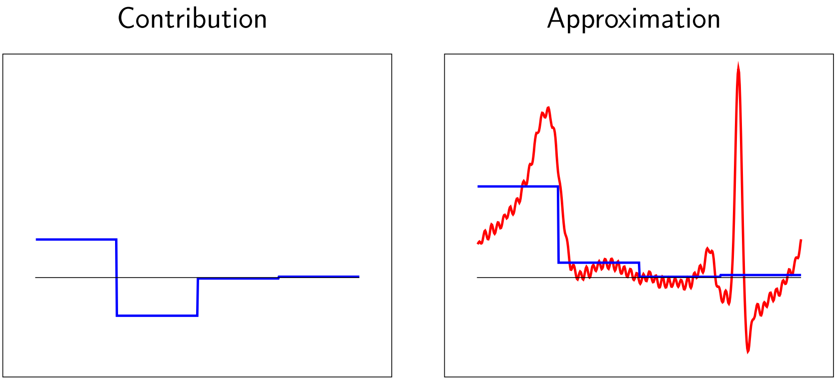
Scale 2^9



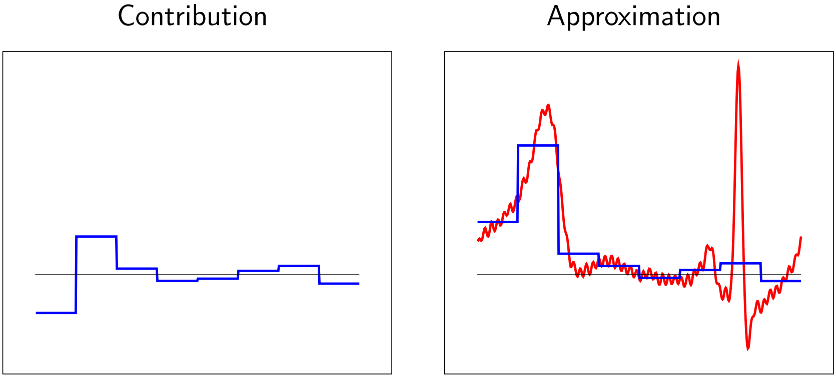
Scale 2^8



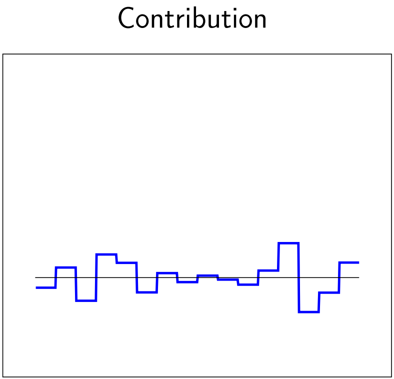
Scale 2^7

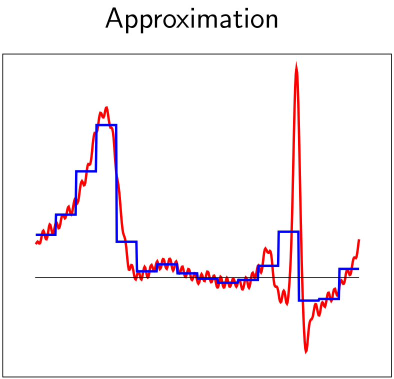


Scale 2^6



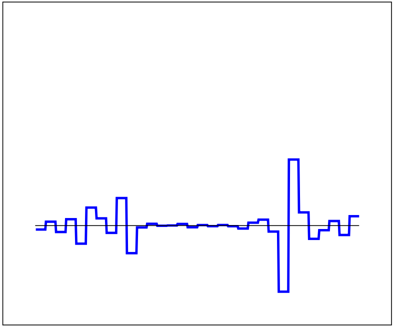
Scale 2^5



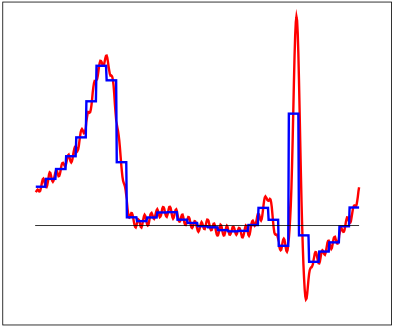


Scale 2^4

Contribution

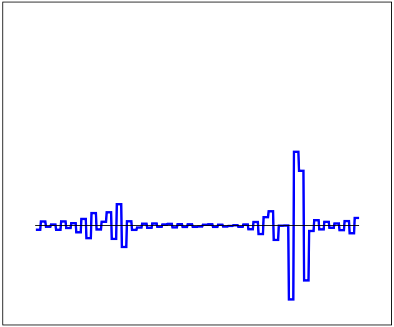


Approximation

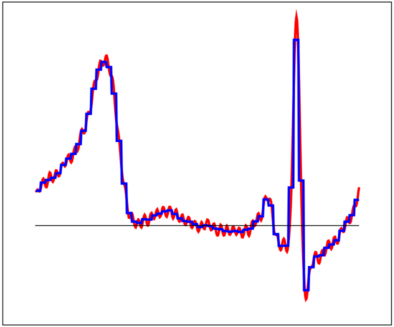


Scale 2^3

Contribution

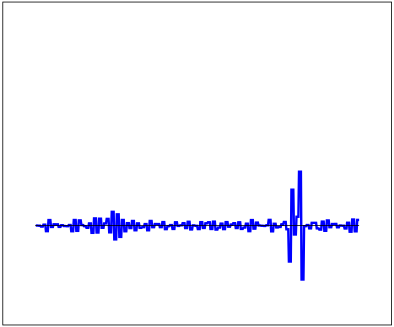


Approximation

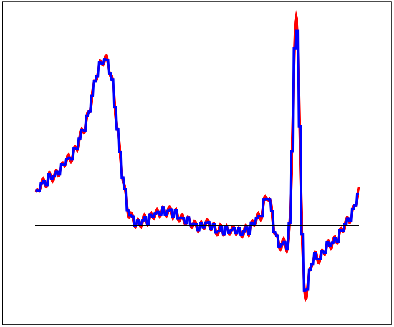


Scale 2^2

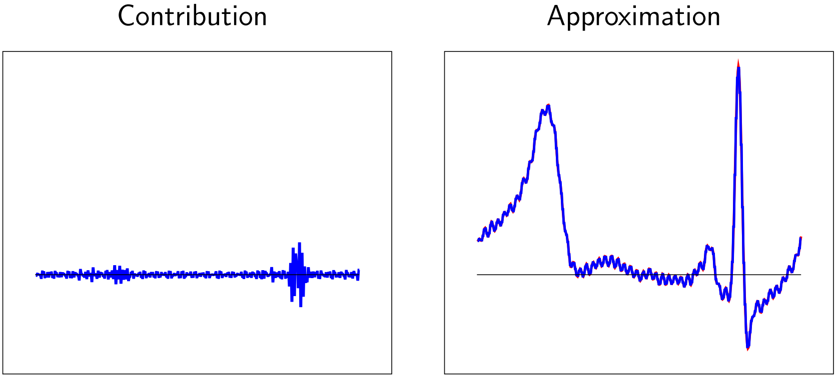
Contribution



Approximation

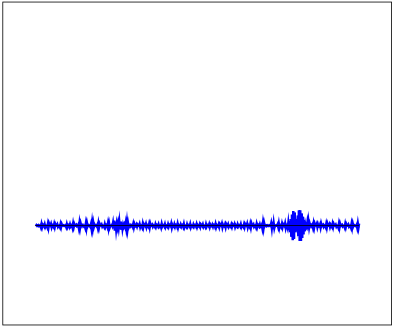


Scale 2^1

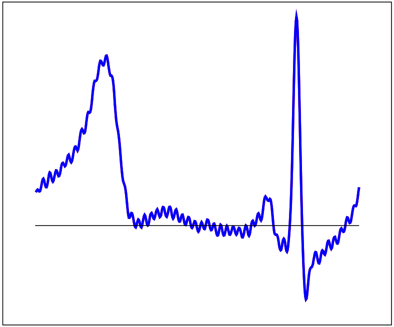


Scale 2^0

Contribution

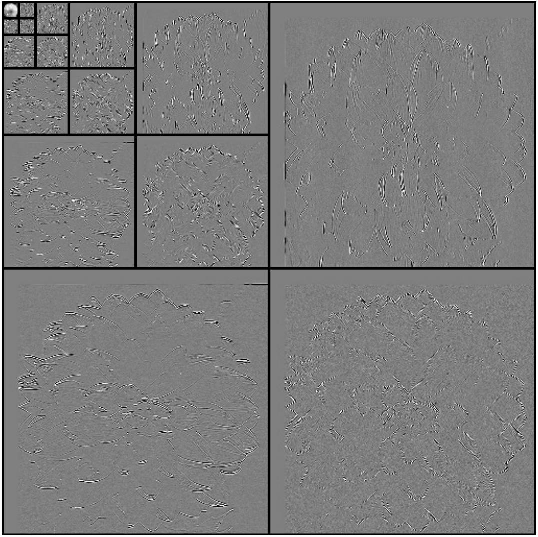


Approximation

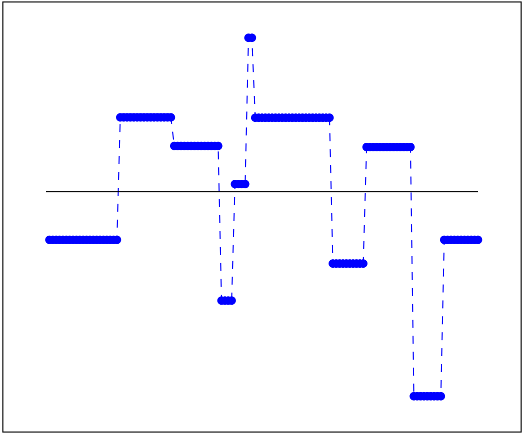


2D wavelet transform

2D wavelet transform



Finite differences



Learning signal representations

Aim: Learn representation from a set of n signals

$$X := \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$$

For each signal

$$x_j \approx \sum_{i=1}^{k} \Phi_i \, A_{ij}, \quad 1 \leq j \leq n, \quad \text{for } k \ll n$$

- $\Phi_1, \ldots, \Phi_k \in \mathbb{R}^d$ are atoms
- $A_1, \ldots, A_n \in \mathbb{R}^k$ are coefficient vectors

Learning signal representations

Equivalent formulation

$$X \approx \begin{bmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_k \end{bmatrix} \begin{bmatrix} A_1 & A_2 & \cdots & A_n \end{bmatrix} = \Phi \, A$$

$\Phi \in \mathbb{R}^{d \times k}$, $A \in \mathbb{R}^{k \times n}$

Learning signal representations

- k means
- Principal-component analysis
- Nonnegative matrix factorization
- Sparse principal-component analysis
- Dictionary learning

k means

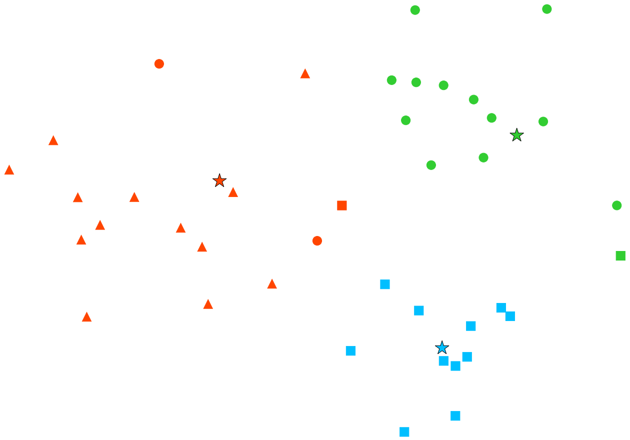
Aim: Divide $x_1, \ldots x_n$ into k classes

Learn $\Phi_1, \ldots, \Phi_k$ that minimize

$$\sum_{i=1}^{n} \left|\left|x_i - \Phi_{c(i)}\right|\right|_2^2$$

$$c\,(i) := \arg \min_{1 \leq j \leq k} \left|\left|x_i - \Phi_j\right|\right|_2$$

k means



Principal-component analysis

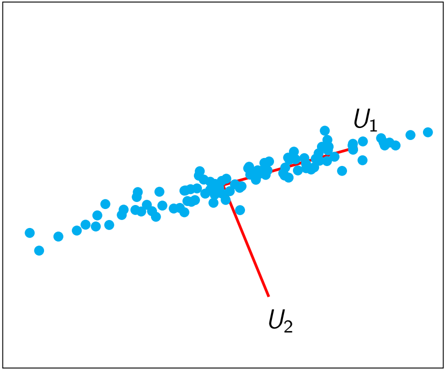
Best rank-k approximation

$$\Phi\, A = U_{1:k} \Sigma_{1:k} V_{1:k}^T = \arg \min_{\left\{\tilde{M} \,\mid\, \operatorname{rank}\left(\tilde{M}\right) = k\right\}} \left|\left| X - \tilde{M} \right|\right|_{\mathrm{F}}^2$$

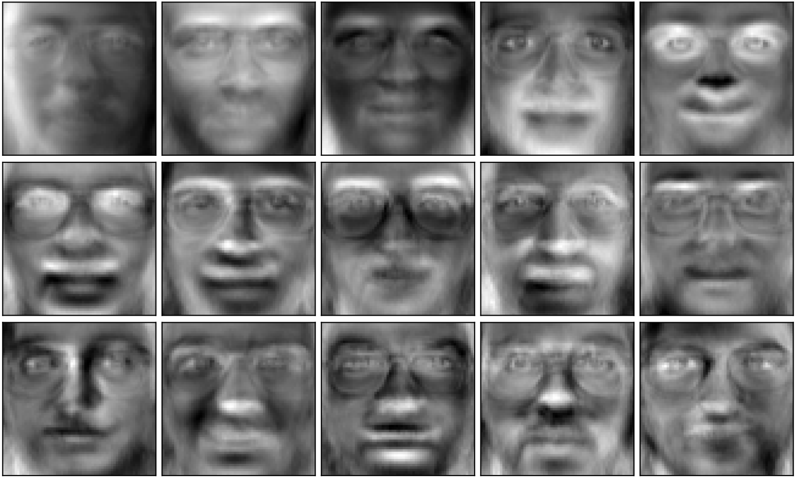
The atoms $\Phi_1, \ldots, \Phi_k$ are orthogonal

Principal-component analysis

$$\frac{\sigma_1}{\sqrt{n}} = 1.3490 \qquad \frac{\sigma_2}{\sqrt{n}} = 0.1438$$



PCA



Nonnegative matrix factorization

Nonnegative atoms/coefficients

$$X \approx \Phi\, A, \quad \Phi_{i,j} \geq 0,\ A_{i,j} \geq 0,\ \text{for all } i,j$$

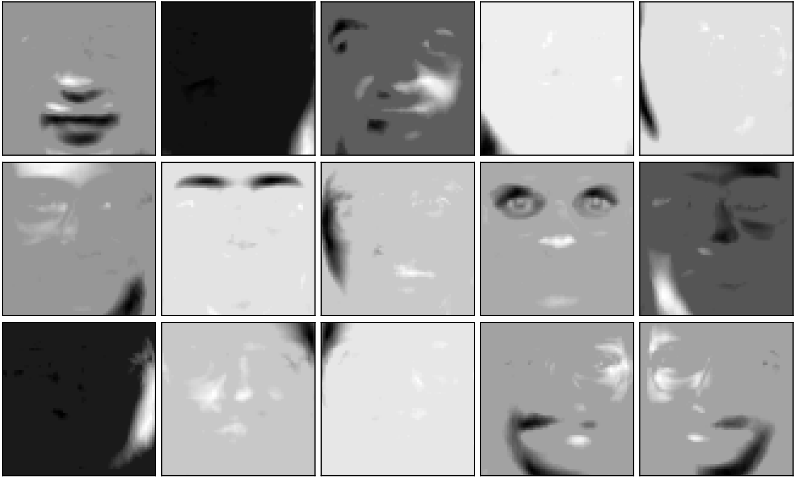
Faces dataset

Sparse PCA

Sparse atoms

$$X \approx \Phi\, A, \quad \Phi \text{ sparse}$$

Faces dataset

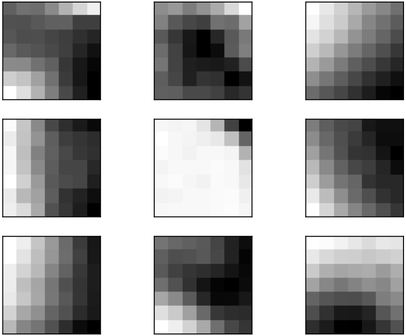


Dictionary learning

Sparse coefficients

$$X \approx \Phi\, A, \quad A \text{ sparse}$$

Dictionary learning



Data-analysis problems

Signal structure

Methods

- General techniques
- Denoising
- Signal recovery
- Signal separation
- Regression
- Compression / dimensionality reduction
- Clustering

When is the problem well posed?

Theoretical analysis

Promoting sparsity

Find sparse x such that $x \approx y$

Hard thresholding

$$\mathcal{H}_\eta\left(y\right)_i := \begin{cases} y_i & \text{if } \left|y_i\right| > \eta \\ 0 & \text{otherwise} \end{cases}$$

Promoting group sparsity

Find group sparse x such that $x \approx y$

Block thresholding

$$\mathcal{B}_{\eta}\left(x\right)_i := \begin{cases} x_i & \text{if } i \in \mathcal{G}_j \quad \text{such that } \left|\left|x_{\mathcal{G}_j}\right|\right|_2 > \eta \\ 0 & \text{otherwise} \end{cases}$$

Promoting low-rank structure

Find low rank M such that $M \approx Y \in \mathbb{R}^{m \times n}$

- Truncate singular-value decomposition $Y = U \Sigma V^T$

$$M = U_{1:k} \Sigma_{1:k} V_{1:k}^T$$

  Solves PCA problem

- Fit $M = A\,B$, $A \in \mathbb{R}^{m \times k}$, $B \in \mathbb{R}^{k \times n}$, by solving

$$\text{minimize} \quad \left|\left| Y - \tilde{A}\,\tilde{B} \right|\right|_{\mathrm{F}}$$

Promoting additional structure in low-rank models

- Nonnegative factors

$$\text{minimize} \quad \left|\left|Y - \tilde{A}\,\tilde{B}\right|\right|_{\mathrm{F}} \quad \text{subject to} \quad \tilde{A}_{i,j} \geq 0$$
$$\tilde{B}_{i,j} \geq 0 \qquad \text{for all } i, j$$

- Sparse factors

$$\text{minimize} \quad \left|\left|Y - \tilde{A}\,\tilde{B}\right|\right|_{\mathrm{F}} + \lambda \sum_{i=1}^{k} \left|\left|\tilde{A}_i\right|\right|_1$$
$$\text{subject to} \quad \left|\left|\tilde{A}_i\right|\right|_2 = 1, \qquad 1 \leq i \leq k$$

$$\text{minimize} \quad \left|\left|Y - \tilde{A}\,\tilde{B}\right|\right|_{\mathrm{F}} + \lambda \sum_{i=1}^{k} \left|\left|\tilde{B}_i\right|\right|_1$$
$$\text{subject to} \quad \left|\left|\tilde{A}_i\right|\right|_2 = 1, \qquad 1 \leq i \leq k$$

Linear models

Signal representation

$$x = D\,c$$

- Columns of D are designed/learned atoms

Linear models

Inverse problems

$$y = A\,x$$

- A models the measurement process

Linear models

Linear regression

$$y = X\ \beta$$

- X contains the predictors

Least squares

Find x such that $Ax \approx y$

$$\text{minimize} \quad ||y - A\tilde{x}||_2$$

Alternatives: Logistic loss for classification

Promoting sparsity

Find sparse x such that $Ax \approx y$

- Greedy methods: Choose entries of x sequentially to minimize residual (matching pursuit, orthogonal m. p., forward stepwise regression)
- Penalize ℓ_1 norm of x

$$\text{minimize} \quad ||y - A\tilde{x}||_2^2 + \lambda||\tilde{x}||_1$$

Implementation:

gradient descent + soft-thresholding / coordinate descent

Promoting group sparsity

Find group sparse x such that $Ax \approx y$

- Penalize ℓ_1/ℓ_2 norm of x

$$\text{minimize} \quad ||y - A\tilde{x}||_2^2 + \lambda ||\tilde{x}||_{1,2}$$

Implementation:

gradient descent + block soft-thresholding / coordinate descent

Promoting low-rank structure

Find low rank M such that $M_\Omega \approx Y_\Omega \in \mathbb{R}^{m\times n}$ for a set of entries Ω

- Penalize nuclear norm of x

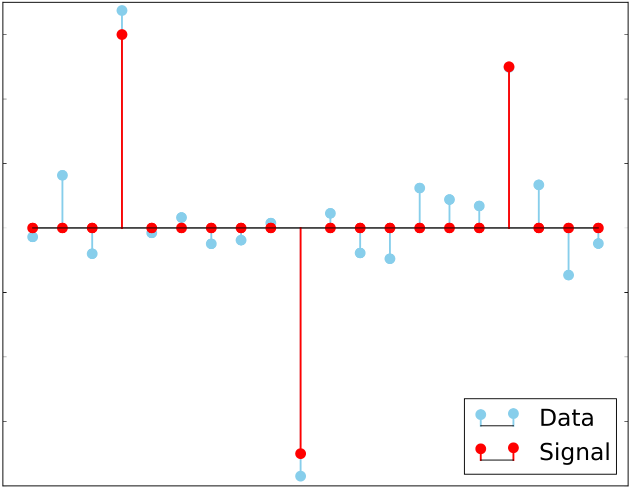
$$\text{minimize} \quad \left|\left| Y_\Omega - \tilde{M}_\Omega \right|\right|_2^2 + \lambda \left|\left| \tilde{M} \right|\right|_*$$

Implementation: gradient descent + soft-thresholding of singular values

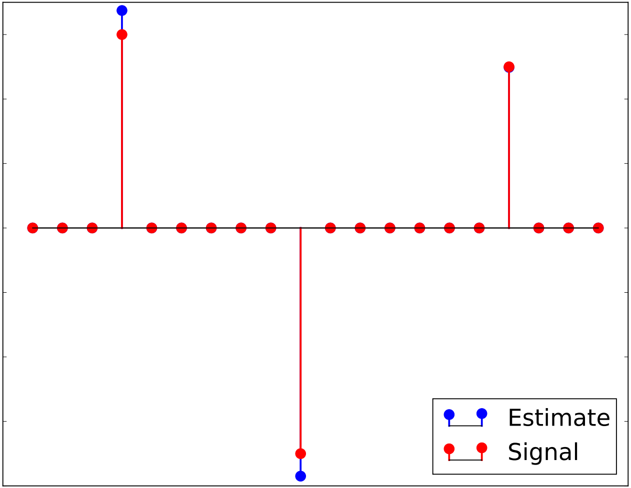
- Fit $M = A\,B$, $A \in \mathbb{R}^{m\times k}$, $B \in \mathbb{R}^{k\times n}$, by solving

$$\text{minimize} \quad \left|\left| Y_\Omega - \left(\tilde{A}\,\tilde{B}\right)_\Omega \right|\right|_{\text{F}}$$

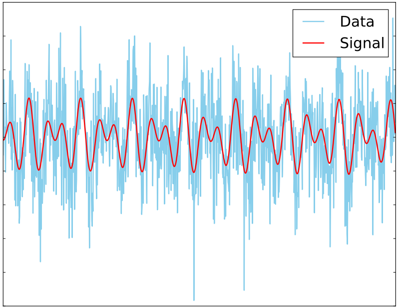
Denoising



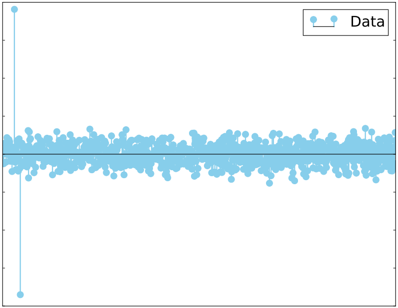
Denoising via thresholding



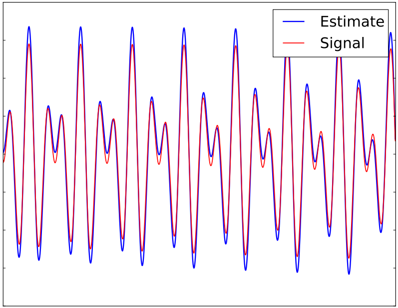
Denoising

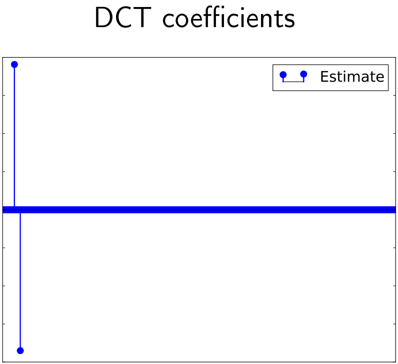


DCT coefficients



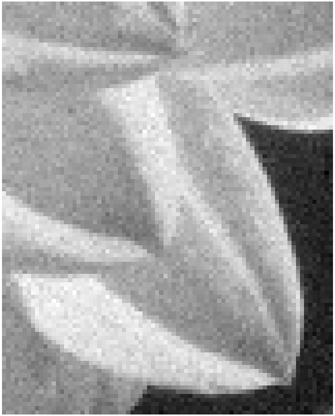
Denoising via thresholding in DCT basis



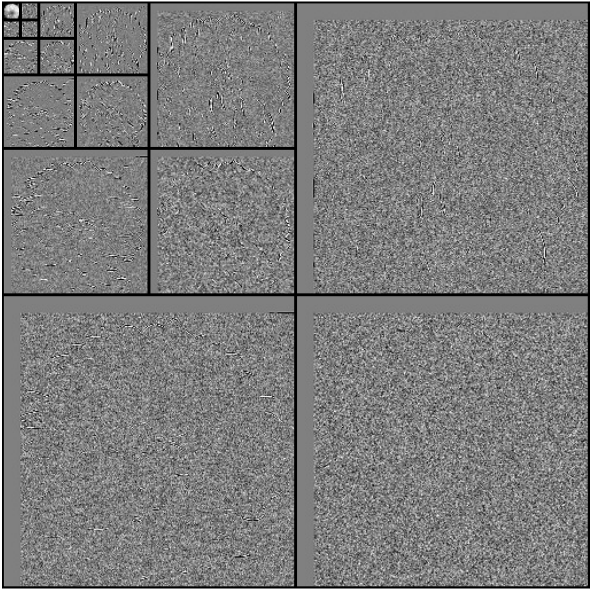


Denoising

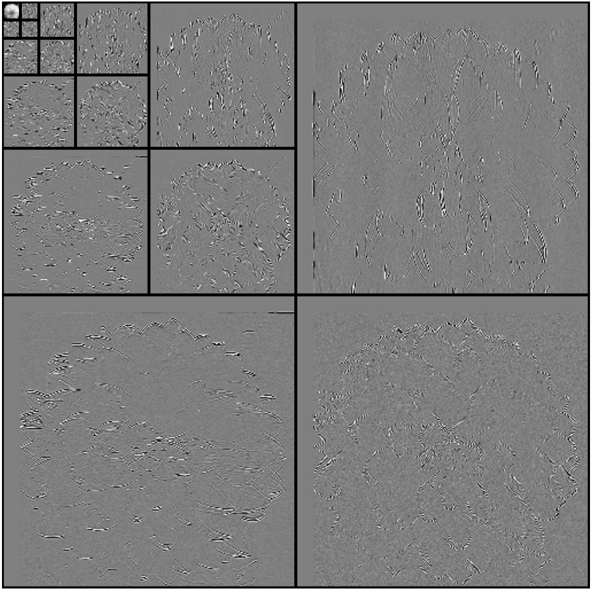
Denoising



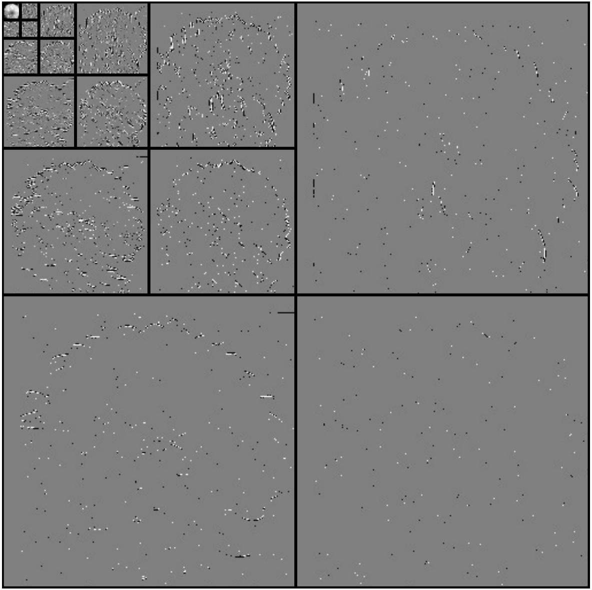
2D wavelet coefficients



Original coefficients



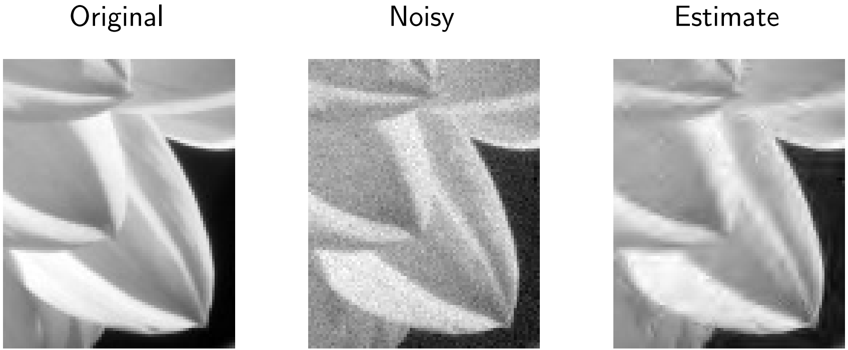
Thresholded coefficients



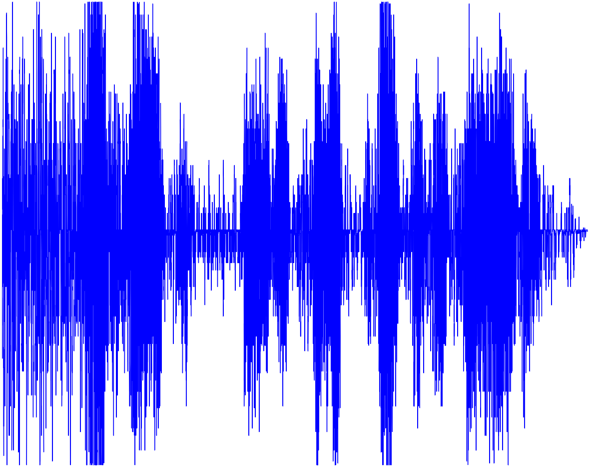
Denoising via thresholding in a wavelet basis

Denoising via thresholding in a wavelet basis

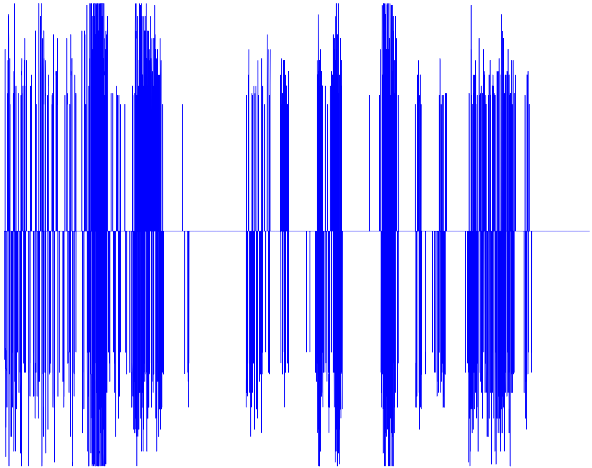
Denoising via thresholding in a wavelet basis



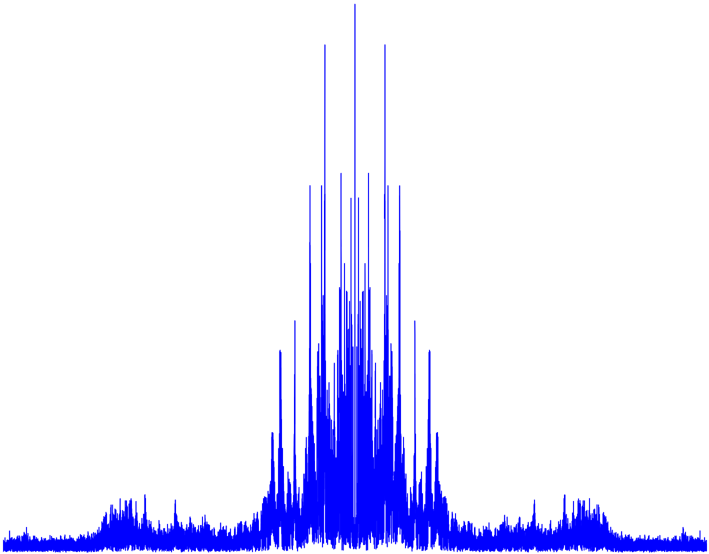
Speech denoising



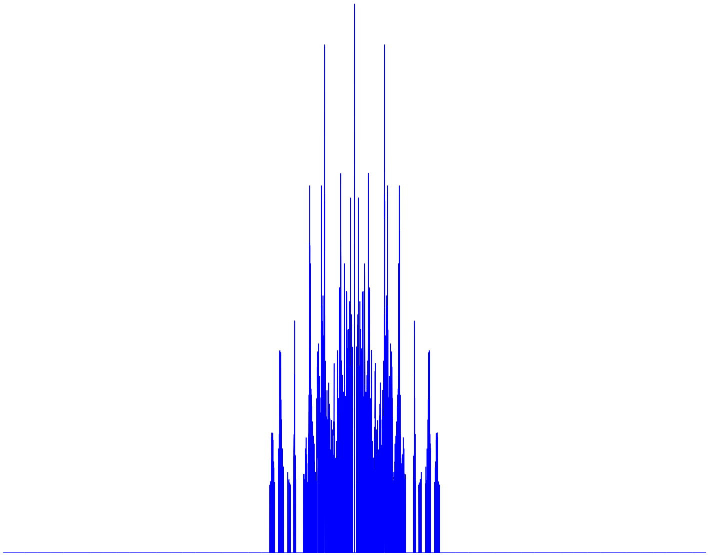
Time thresholding



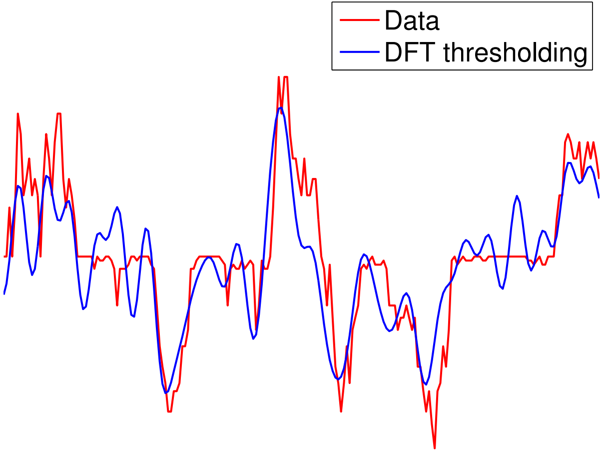
Spectrum



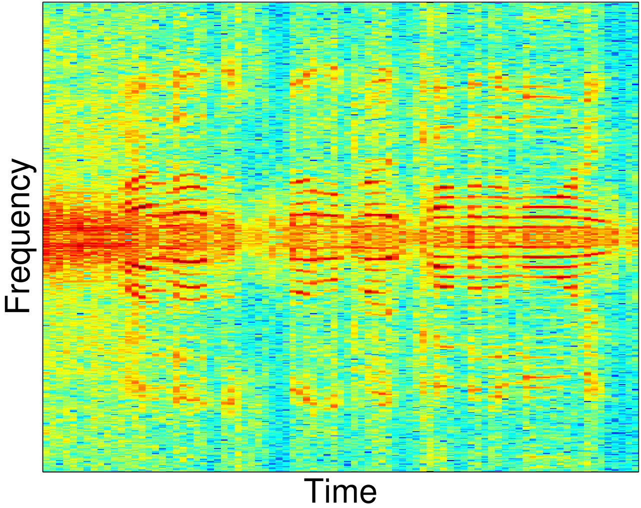
Frequency thresholding



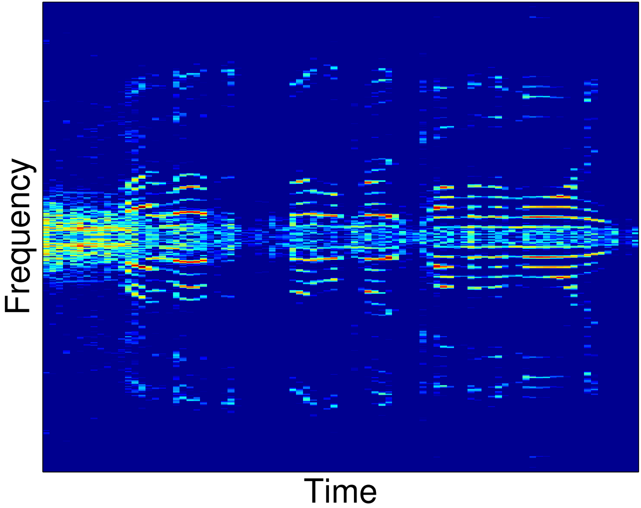
Frequency thresholding



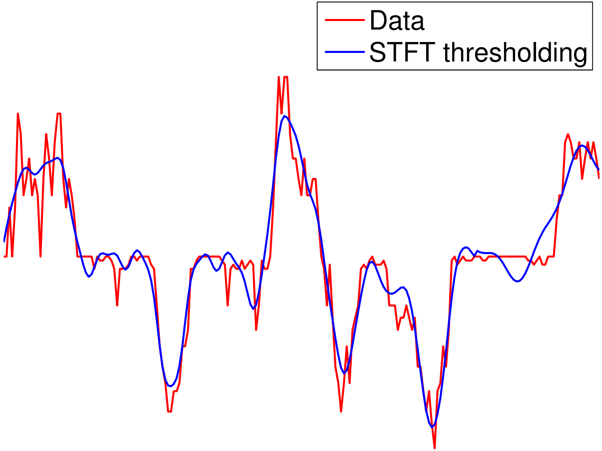
Spectrogram (STFT)



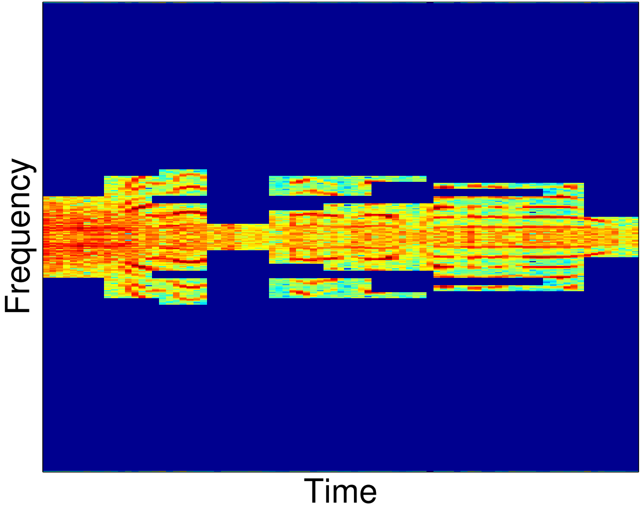
STFT thresholding



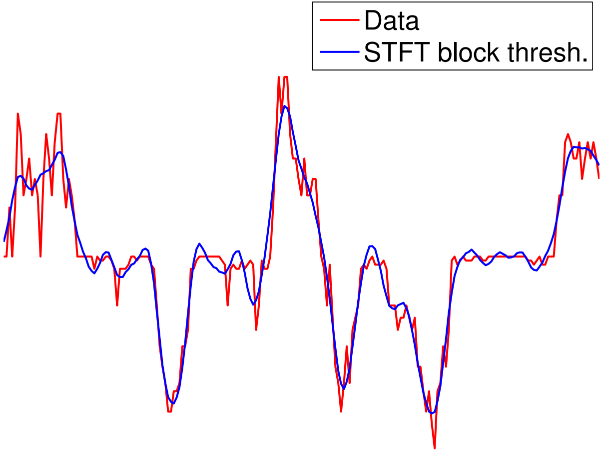
STFT thresholding



STFT block thresholding

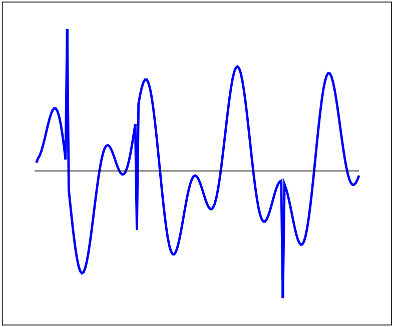


STFT block thresholding

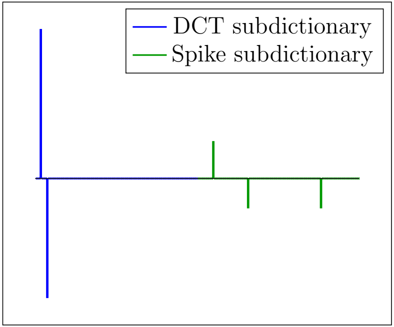


Sines and spikes

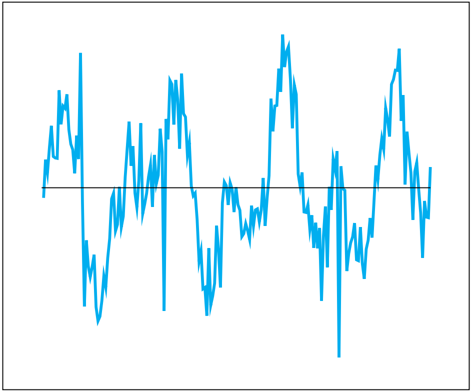
$x = Dc$



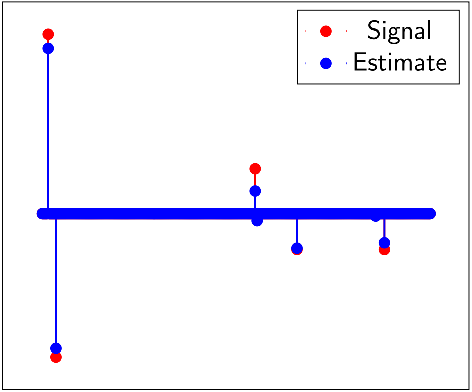
c



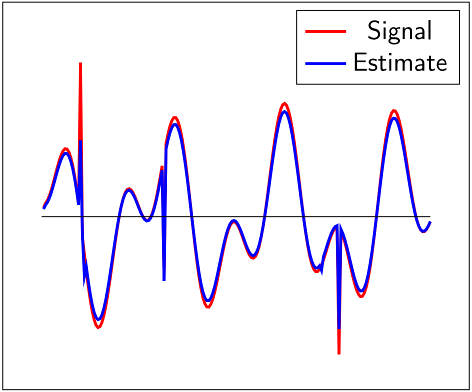
Denoising



Denoising via ℓ_1-norm regularized least squares

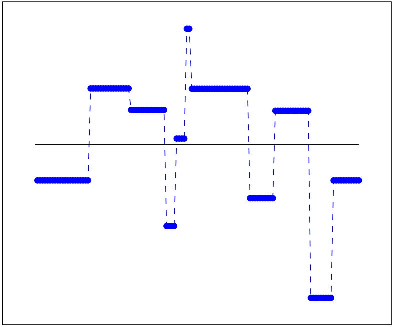


Denoising via ℓ_1-norm regularized least squares

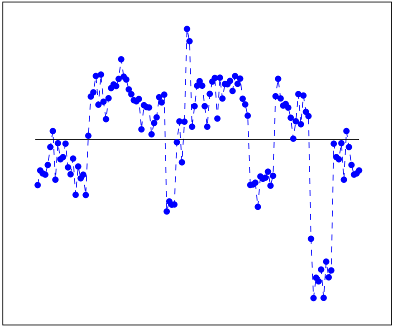


Denoising

Signal

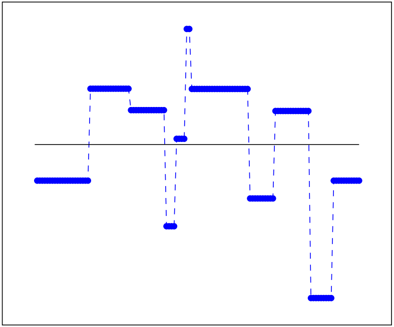


Data

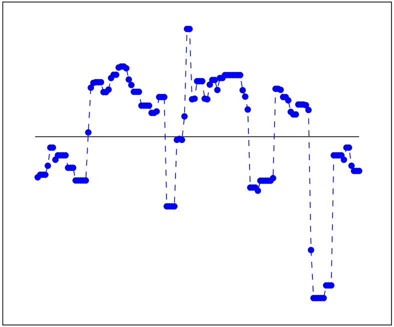


Denoising via TV regularization

Signal

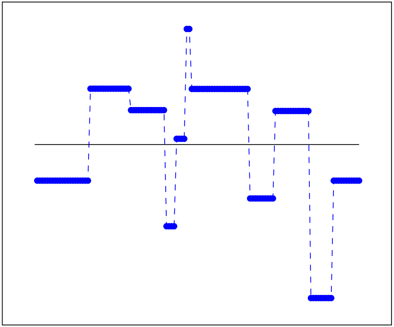


TV reg. (small λ)

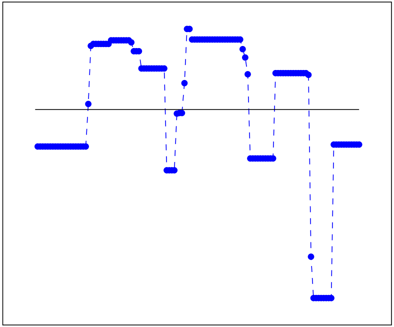


Denoising via TV regularization

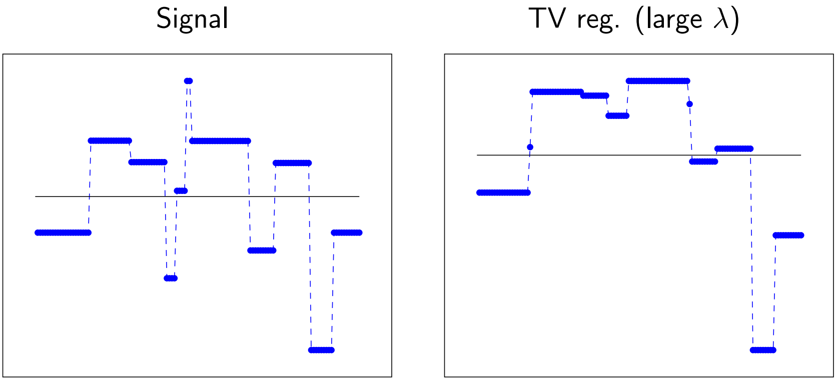
Signal



TV reg. (medium λ)



Denoising via TV regularization



Denoising via TV regularization

Denoising via TV regularization

Small λ

Small λ

Medium λ

Medium λ

Large λ

Large λ

Denoising via TV regularization

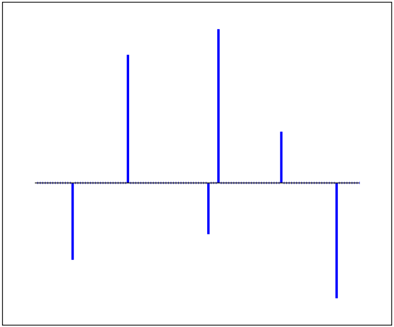
Original

Noisy

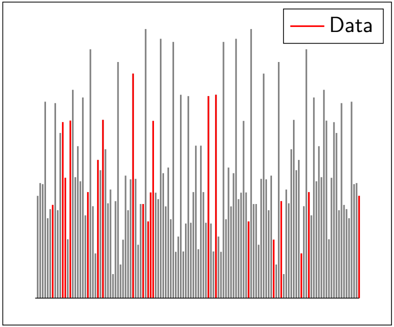
Estimate

Compressed sensing

Signal

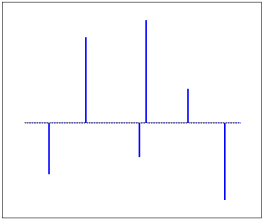


Spectrum

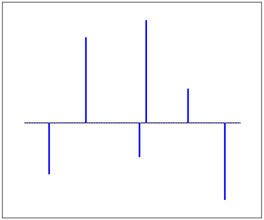


ℓ_1-norm minimization

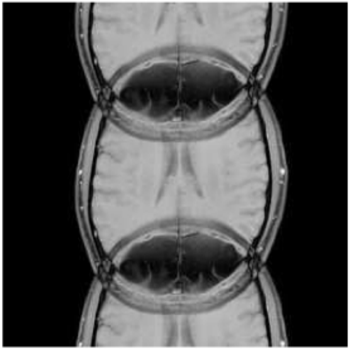
Signal

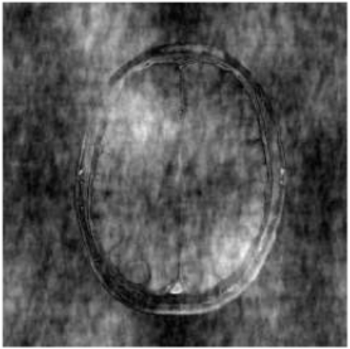


Estimate



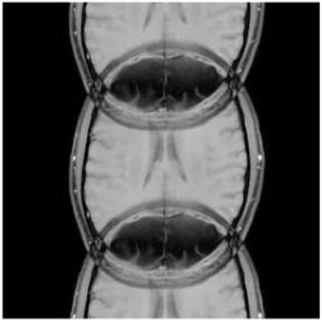
x2 undersampling



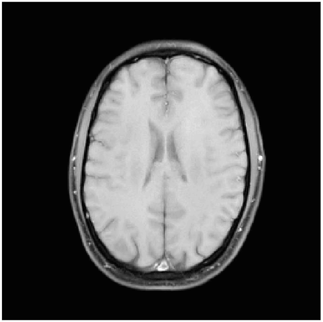


ℓ_1-norm minimization

Regular

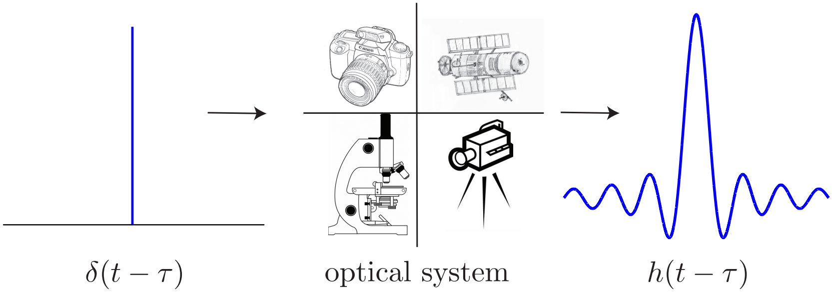


Random



Super-resolution

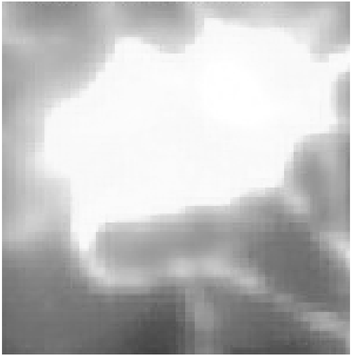
The resolving power of lenses, however perfect, is limited (**Lord Rayleigh**)

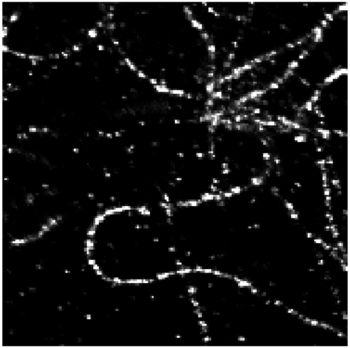


Diffraction imposes a fundamental limit on the resolution of optical systems

Super-resolution

Super-resolution: ℓ_1-norm regularization



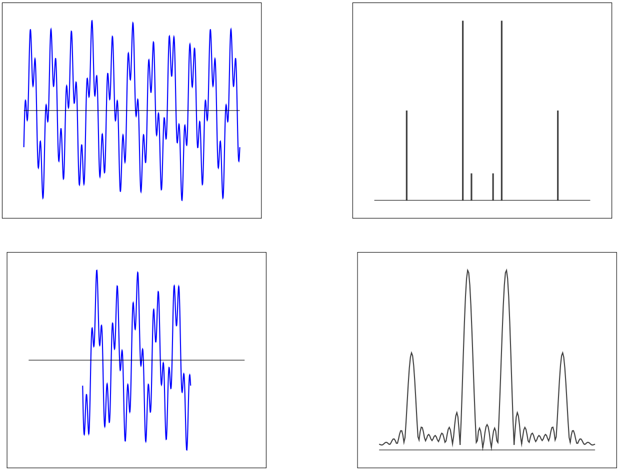


Spectral Super-resolution

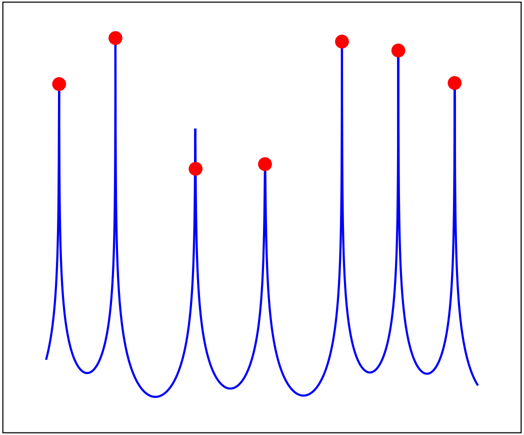
Spectrum

Signal

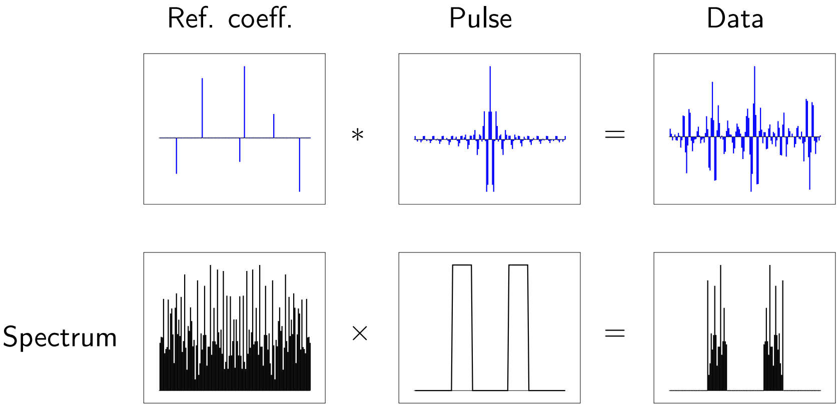
Data



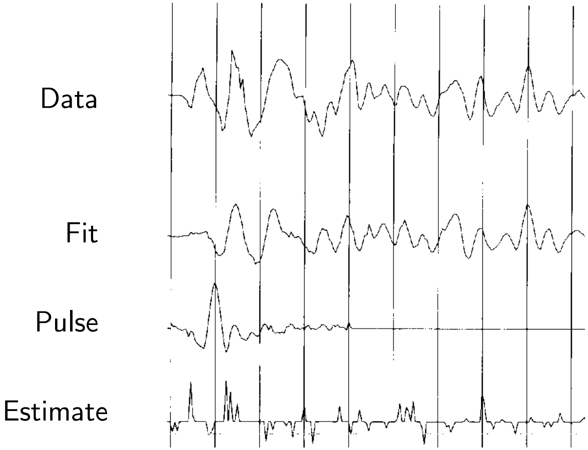
Spectral super-resolution: Pseudospectrum from low-rank model (MUSIC)



Deconvolution



Deconvolution with the ℓ_1 norm (Taylor, Banks, McCoy '79)



Matrix completion

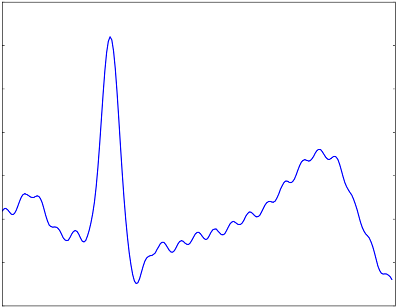
$$
\begin{array}{c l}
\begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} & \\
\begin{pmatrix}
1 & ? & 5 & 4 \\
? & 1 & 4 & 5 \\
4 & 5 & 2 & ? \\
5 & 4 & 2 & 1 \\
4 & 5 & 1 & 2 \\
1 & 2 & ? & 5
\end{pmatrix} &
\begin{array}{l}
\text{The Dark Knight} \\
\text{Spiderman 3} \\
\text{Love Actually} \\
\text{Bridget Jones's Diary} \\
\text{Pretty Woman} \\
\text{Superman 2}
\end{array}
\end{array}
$$

Matrix completion via nuclear-norm minimization

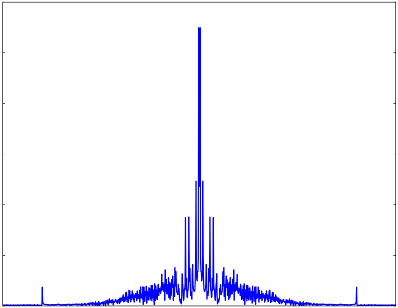
$$
\begin{array}{cccccl}
 & \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} & \\
\left(\begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array}\right. & & & & & \\
\end{array}
$$

$$
\begin{array}{l}
\begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} \\
\left(\begin{array}{cccc}
1 & {\color{red}2}\ (1) & 5 & 4 \\
{\color{red}2}\ (2) & 1 & 4 & 5 \\
4 & 5 & 2 & {\color{red}2}\ (1) \\
5 & 4 & 2 & 1 \\
4 & 5 & 1 & 2 \\
1 & 2 & {\color{red}5}\ (5) & 5
\end{array}\right)
\begin{array}{l}
\text{The Dark Knight} \\
\text{Spiderman 3} \\
\text{Love Actually} \\
\text{Bridget Jones's Diary} \\
\text{Pretty Woman} \\
\text{Superman 2}
\end{array}
\end{array}
$$

Electrocardiogram

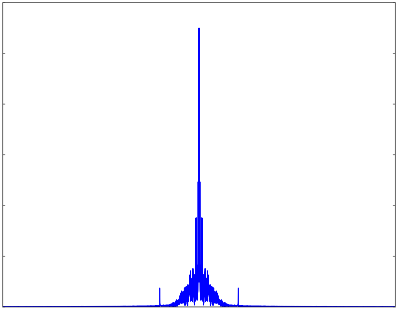


Spectrum

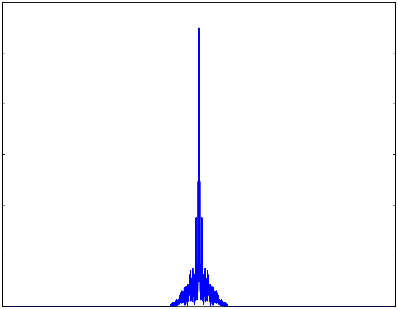


Electrocardiogram: High-frequency noise (power line hum)

Original spectrum

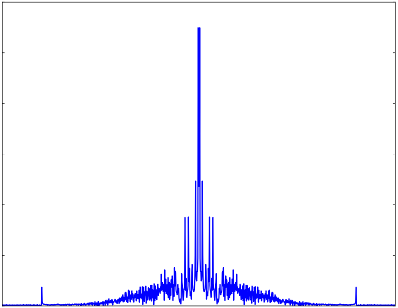


Low-pass filtered spectrum

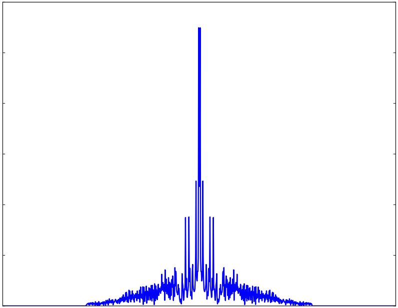


Electrocardiogram: High-frequency noise (power line hum)

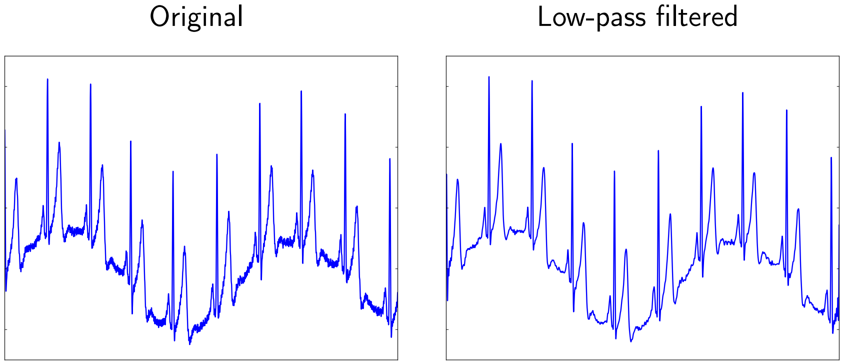
Original spectrum



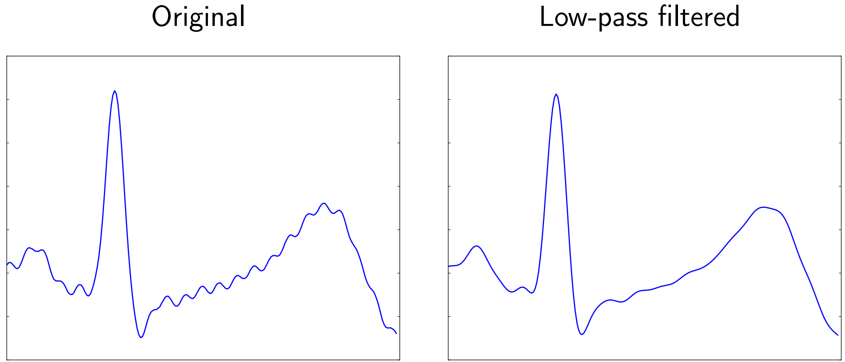
Low-pass filtered spectrum



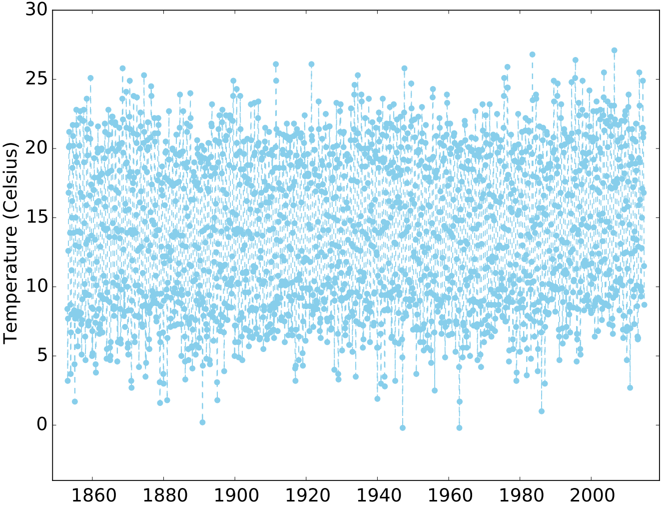
Electrocardiogram: High-frequency noise (power line hum)



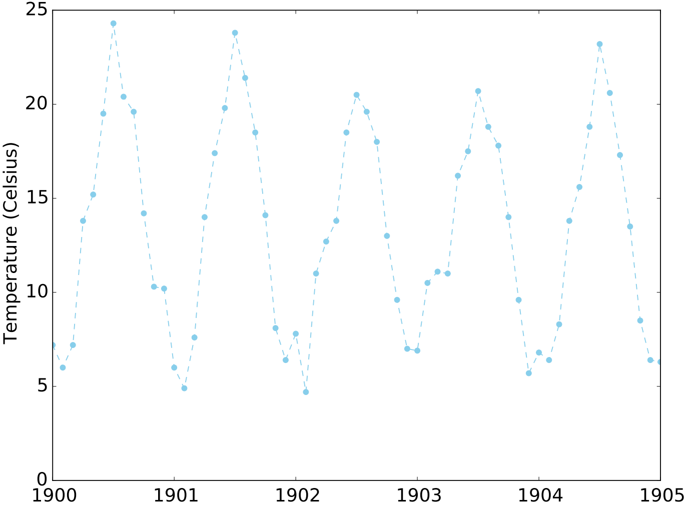
Electrocardiogram: High-frequency noise (power line hum)



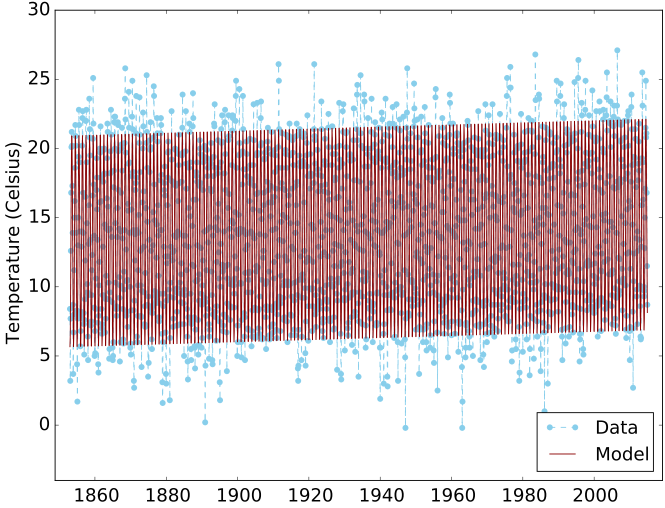
Temperature data



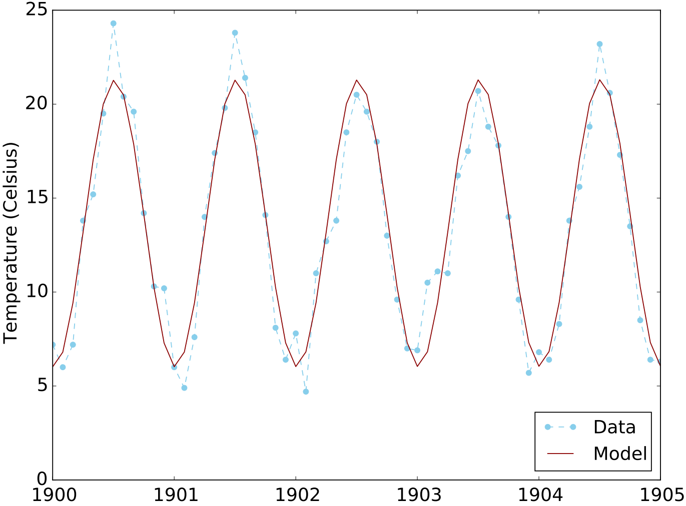
Temperature data



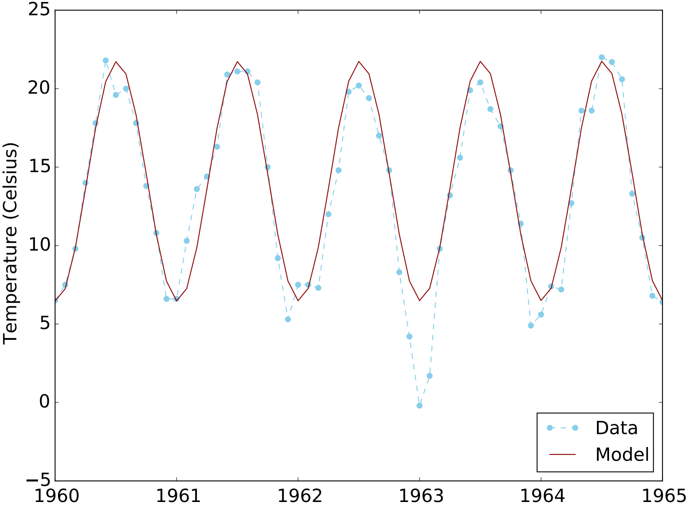
Model fitted by least squares



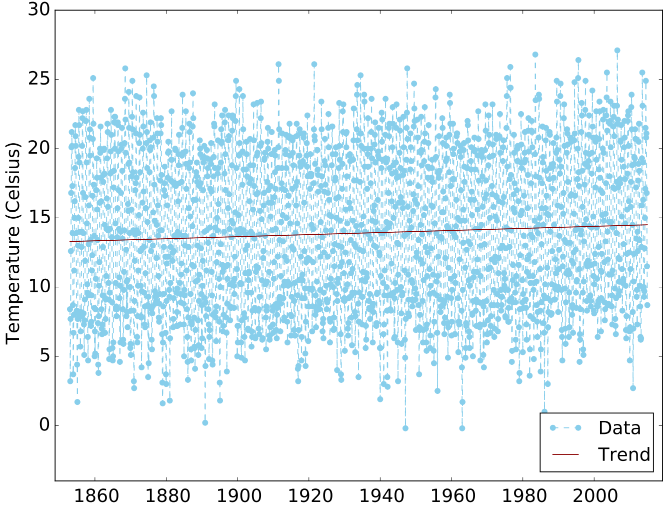
Model fitted by least squares



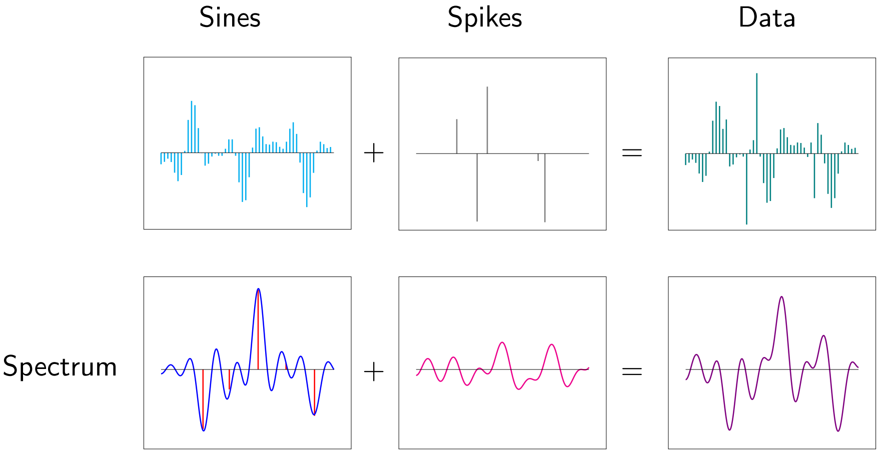
Model fitted by least squares



Trend: Increase of 0.75 °C / 100 years (1.35 °F)

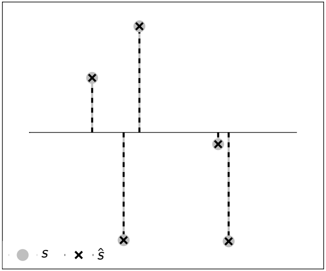


Demixing of sines and spikes

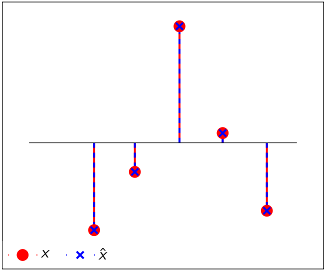


$$\mathcal{F}_c\, x \quad + \quad s \quad = \quad y$$

Demixing of sines and spikes



Spikes



Sines (spectrum)

Background subtraction

Low-rank component

Sparse component

Data-analysis problems

Signal structure

Methods

- General techniques
- Denoising
- Signal recovery
- Signal separation
- Regression
- Compression / dimensionality reduction
- Clustering

When is the problem well posed?

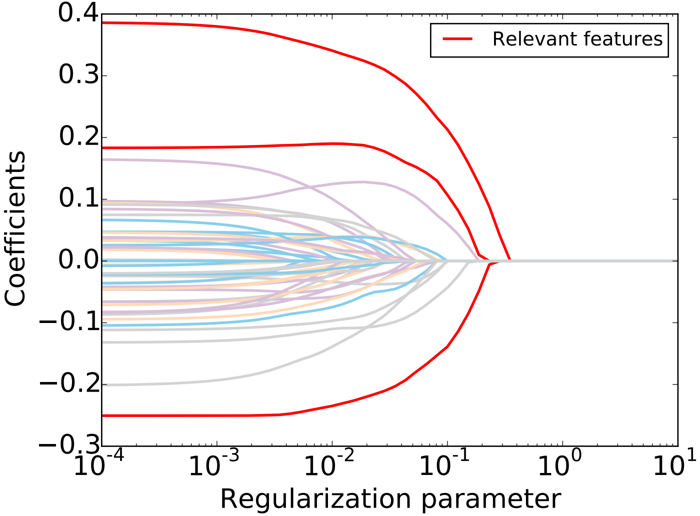
Theoretical analysis

Sparse regression

Assumption: Response only depends on a subset $\mathcal{S}$ of $s \ll p$ predictors

Model-selection problem: Determine what predictors are relevant

Lasso


Relevant features
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
Coefficients
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
Regularization parameter

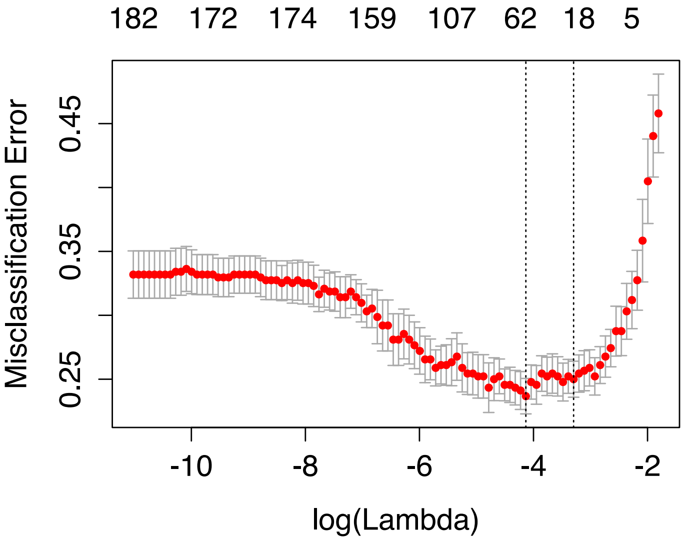
Arrhythmia prediction

Predict whether patient has arrhythmia from $n = 271$ examples and $p = 182$ predictors

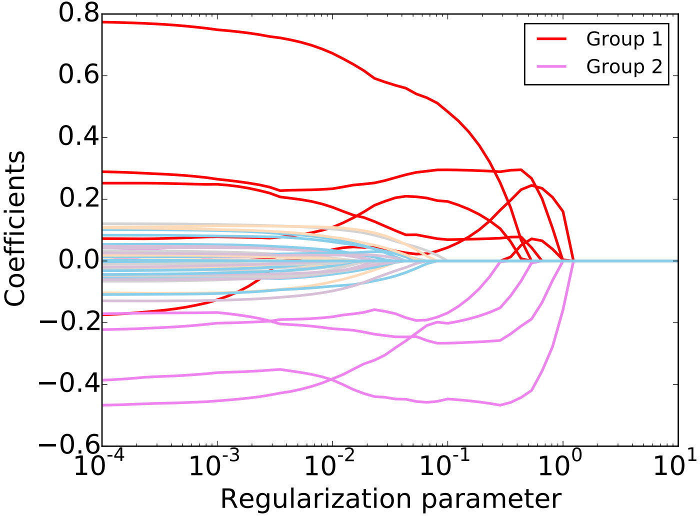
- Age, sex, height, weight
- Features obtained from electrocardiogram recordings

Best sparse model uses around 60 predictors

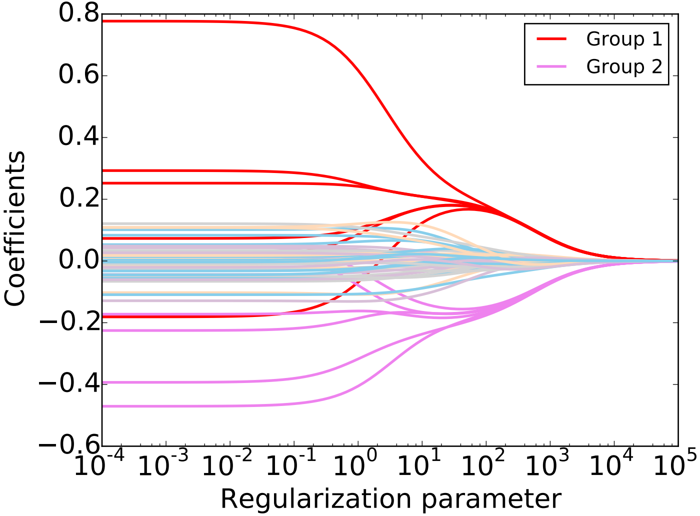
Lasso (logistic regression)



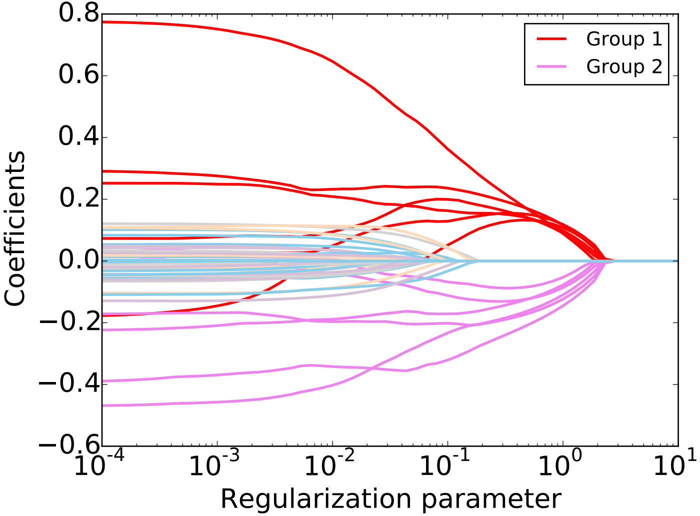
Correlated predictors: Lasso path



Correlated predictors: Ridge-regression path



Correlated predictors: Elastic net path



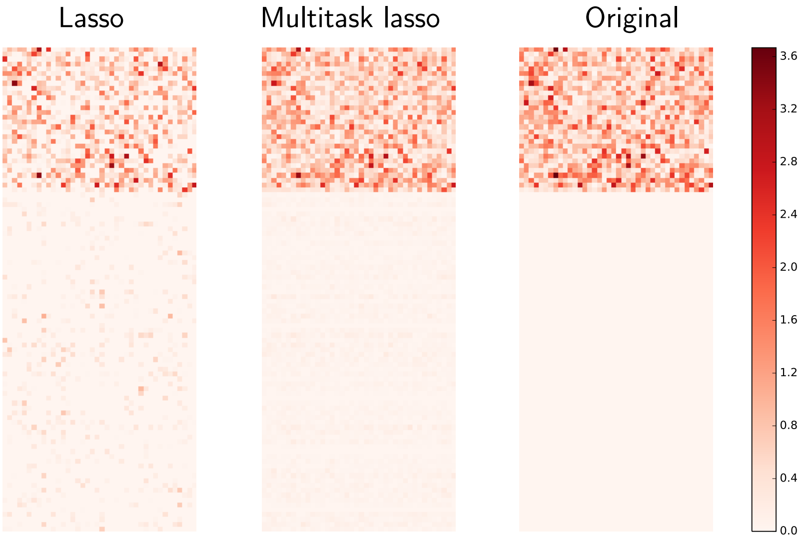
Multi-task learning

Several responses y_1, y_2, ..., y_k modeled with the same predictors

Assumption: Responses depend on the same subset of predictors

Aim: Learn a group-sparse model

Multitask learning



Compression

Original

Compression via frequency representation

10 % largest DCT coeffs

Compression via frequency representation

2 % largest DCT coeffs

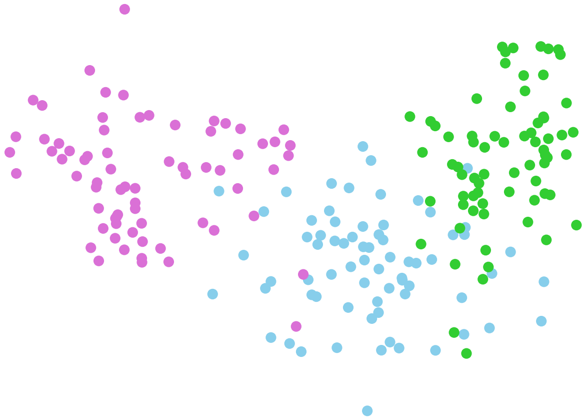
Dimensionality reduction

Seeds from three different varieties of wheat: Kama, Rosa and Canadian

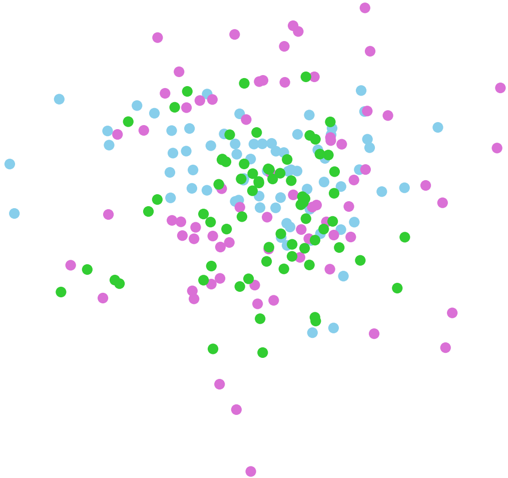
Dimensions:

- Area
- Perimeter
- Compactness
- Length of kernel
- Width of kernel
- Asymmetry coefficient
- Length of kernel groove

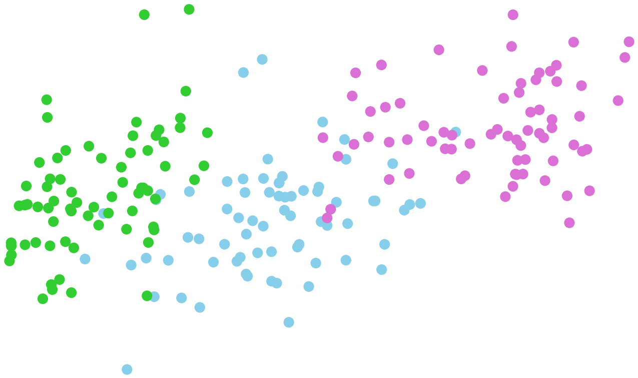
PCA: Projection onto two first PCs



PCA: Projection onto two last PCs



Random projections



Data-analysis problems

Signal structure

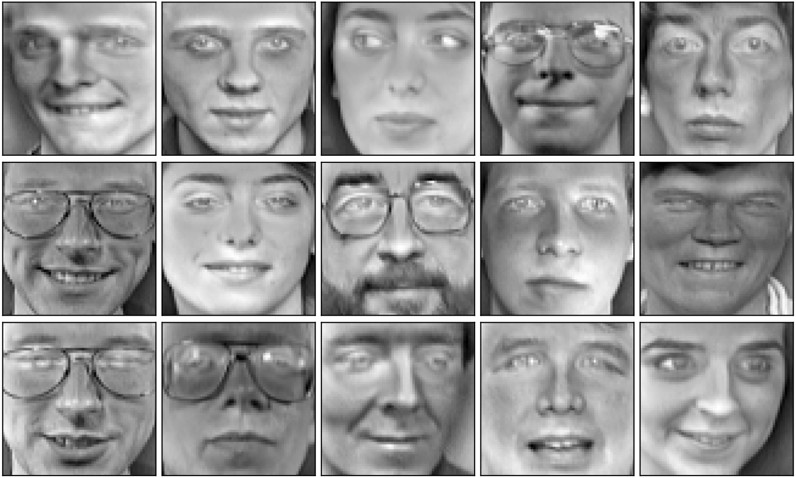
Methods

- General techniques
- Denoising
- Signal recovery
- Signal separation
- Regression
- Compression / dimensionality reduction
- Clustering

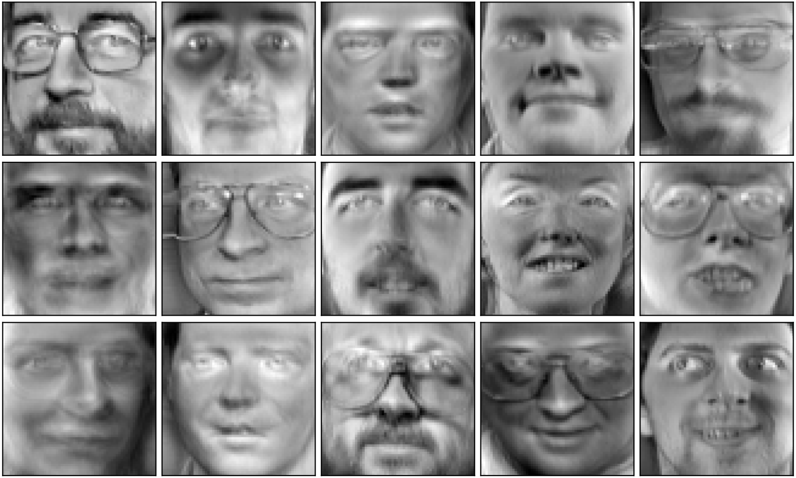
When is the problem well posed?

Theoretical analysis

Clustering



k means using Lloyd's algorithm



Collaborative filtering

$$
\begin{array}{ccccl}
\text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} & \\
\left(\begin{array}{c} 1 \\ 2 \\ 4 \\ 5 \\ 4 \\ 1 \end{array}\right. & \begin{array}{c} 1 \\ 1 \\ 5 \\ 4 \\ 5 \\ 2 \end{array} & \begin{array}{c} 5 \\ 4 \\ 2 \\ 2 \\ 1 \\ 5 \end{array} & \left.\begin{array}{c} 4 \\ 5 \\ 1 \\ 1 \\ 2 \\ 5 \end{array}\right) & \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array}
\end{array}
$$

First left singular vector clusters movies

$$
\begin{array}{cccccccc}
 & & \text{D. Knight} & \text{Sp. 3} & \text{Love Act.} & \text{B.J.'s Diary} & \text{P. Woman} & \text{Sup. 2} \\
U_1 = & (& -0.45 & -0.39 & 0.39 & 0.39 & 0.39 & -0.45 \;)
\end{array}
$$

First right singular vector clusters users

$$\begin{array}{rcccc} & \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \\ V_1 = & (0.48 & 0.52 & -0.48 & -0.52) \end{array}$$

Topic modeling

$$
\begin{array}{c}
\begin{array}{ccccccccc}
\text{singer} & \text{GDP} & \text{senate} & \text{election} & \text{vote} & \text{stock} & \text{bass} & \text{market} & \text{band}
\end{array} & \text{Articles} \\
\begin{pmatrix}
6 & 1 & 1 & 0 & 0 & 1 & 9 & 0 & 8 \\
1 & 0 & 9 & 5 & 8 & 1 & 0 & 1 & 0 \\
8 & 1 & 0 & 1 & 0 & 0 & 9 & 1 & 7 \\
0 & 7 & 1 & 0 & 0 & 9 & 1 & 7 & 0 \\
0 & 5 & 6 & 7 & 5 & 6 & 0 & 7 & 2 \\
1 & 0 & 8 & 5 & 9 & 2 & 0 & 0 & 1
\end{pmatrix} & \begin{array}{c} a \\ b \\ c \\ d \\ e \\ f \end{array}
\end{array}
$$

Right nonnegative factors cluster words

$$
\begin{array}{ccccccccccc}
 & & \text{singer} & \text{GDP} & \text{senate} & \text{election} & \text{vote} & \text{stock} & \text{bass} & \text{market} & \text{band} \\
H_1 & = & (0.34 & 0 & 3.73 & 2.54 & 3.67 & 0.52 & 0 & 0.35 & 0.35) \\
H_2 & = & (\;0 & 2.21 & 0.21 & 0.45 & 0 & 2.64 & 0.21 & 2.43 & 0.22) \\
H_3 & = & (3.22 & 0.37 & 0.19 & 0.2 & 0 & 0.12 & 4.13 & 0.13 & 3.43)
\end{array}
$$

Left nonnegative factors cluster documents

$$
\begin{array}{cccccccc}
 & & a & b & c & d & e & f \\
W_1 & = & (0.03 & 2.23 & 0 & 0 & 1.59 & 2.24) \\
W_2 & = & (\ 0.1 & 0 & 0.08 & 3.13 & 2.32 & 0\ \ ) \\
W_3 & = & (2.13 & 0 & 2.22 & 0 & 0 & 0.03)
\end{array}
$$

Data-analysis problems

Signal structure

Methods

- General techniques
- Denoising
- Signal recovery
- Signal separation
- Regression
- Compression / dimensionality reduction
- Clustering

When is the problem well posed?

Theoretical analysis

Denoising based on sparsity

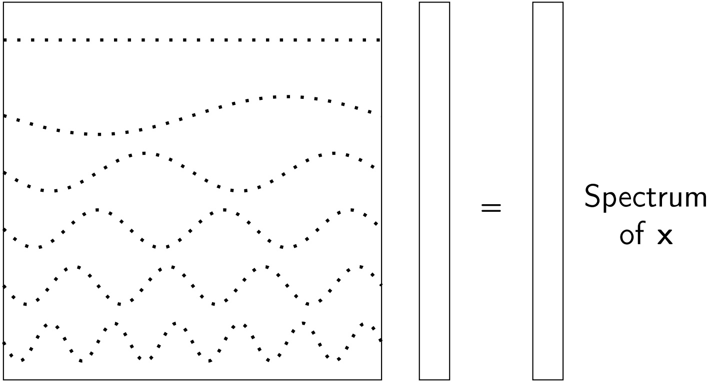
Signal is sparse in the chosen representation

Noise is not sparse in the chosen representation

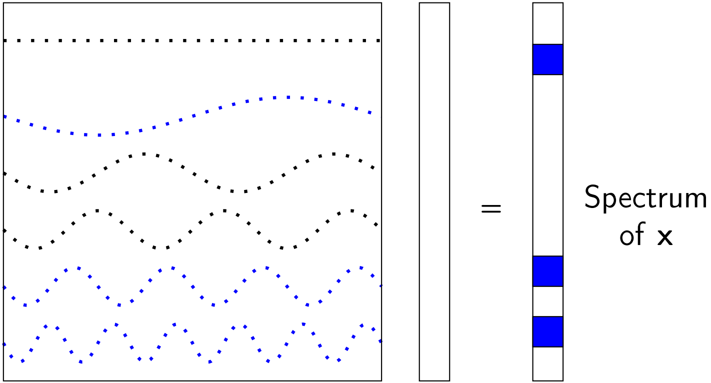
Signal recovery

	Measurements	*Class of signals*
Compressed sensing	Gaussian, random Fourier coeffs.	Sparse
Super-resolution	Low pass	Signals with min. separation
Matrix completion	Random sampling	Incoherent low-rank matrices

Compressed sensing

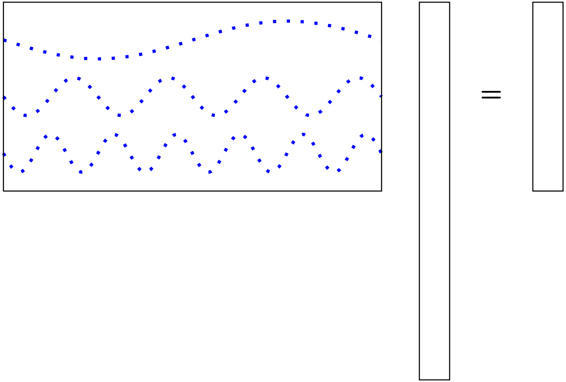


Compressed sensing

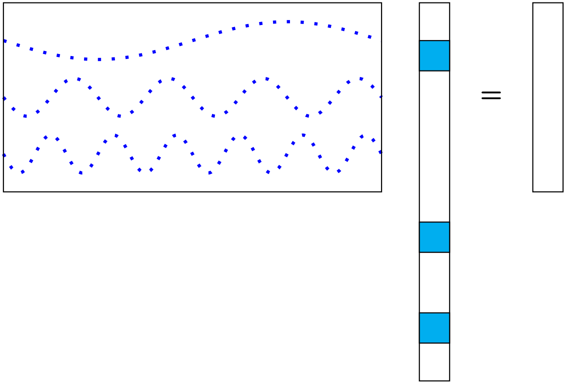


Measurement operator = random frequency samples

Compressed sensing

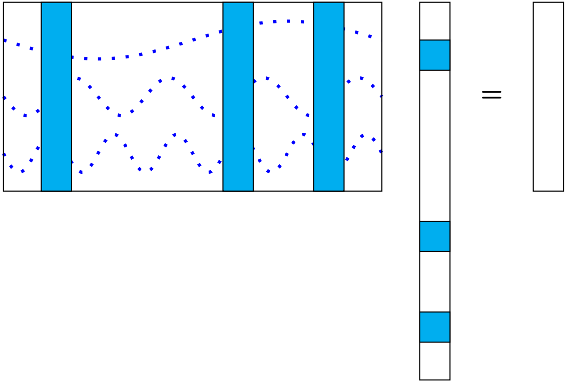


Compressed sensing



Aim: Study effect of measurement operator on sparse vectors

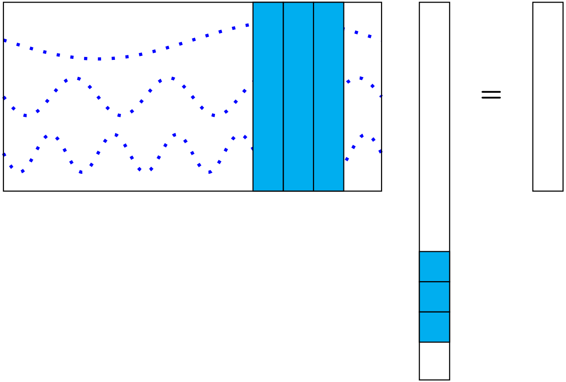
Compressed sensing



Operator is well conditioned when acting upon any sparse signal

(*restricted isometry property*)

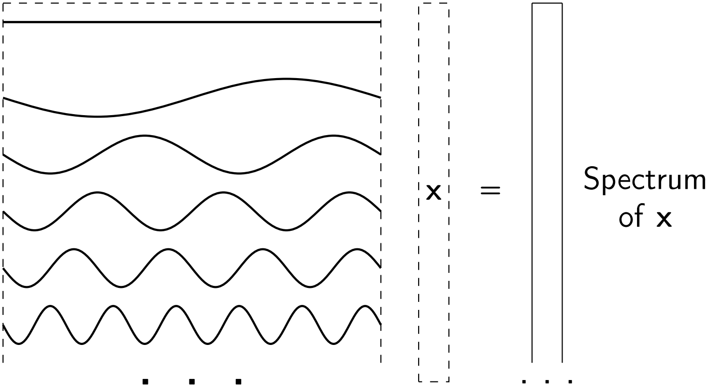
Compressed sensing



Operator is well conditioned when acting upon any sparse signal

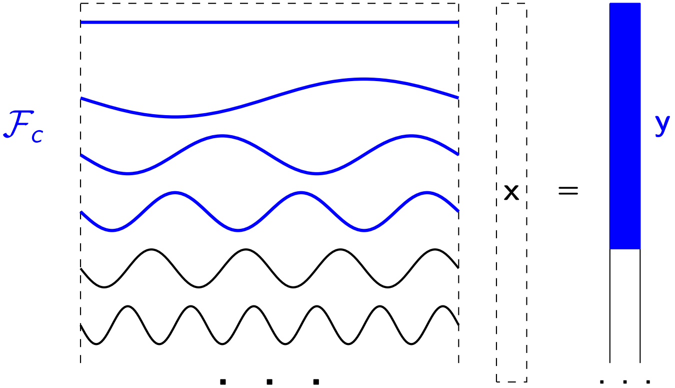
(*restricted isometry property*)

Super-resolution



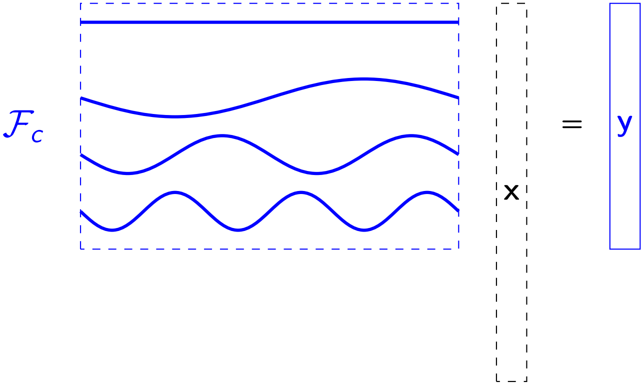
No discretization

Super-resolution



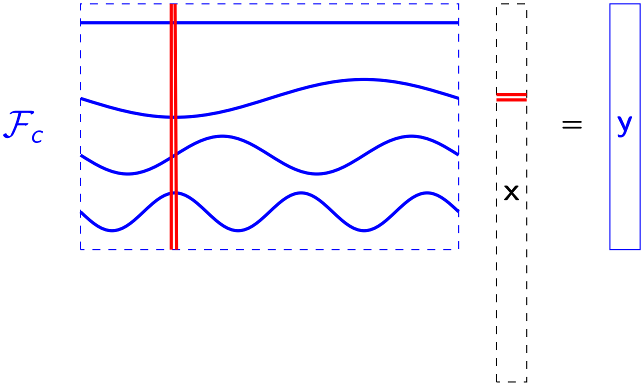
Data: Low-pass Fourier coefficients

Super-resolution



Data: Low-pass Fourier coefficients

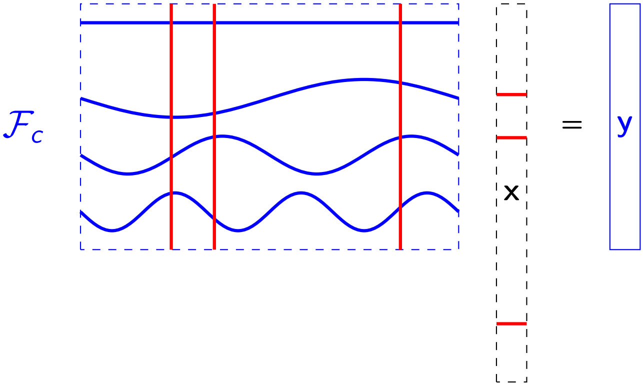
Super-resolution



Problem: If the support is clustered, the problem may be ill posed

In super-resolution sparsity is not enough!

Super-resolution



If the support is spread out, there is still hope

We need conditions beyond sparsity

Matrix completion

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 2 & 3 & 4 & 5 \end{bmatrix}$$

Signal separation

The signals are identifiable

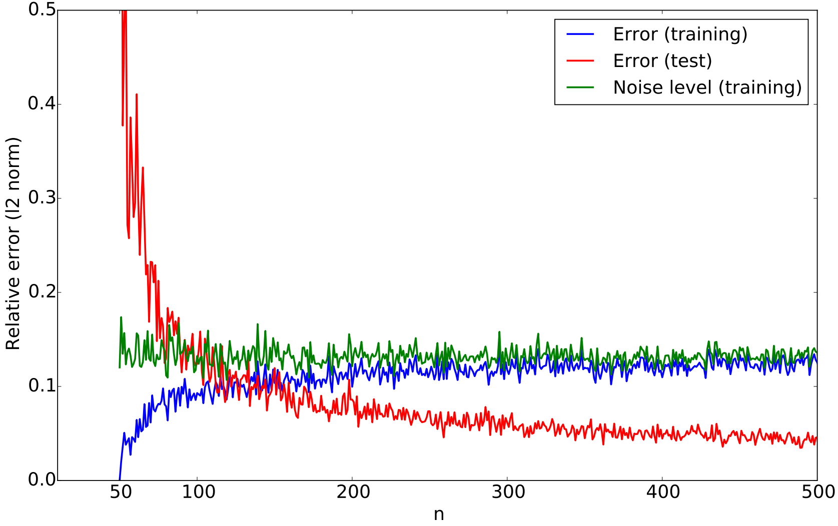
Example: For low rank + sparse model, low rank component cannot be sparse and vice versa

Regression

Enough examples to prevent overfitting

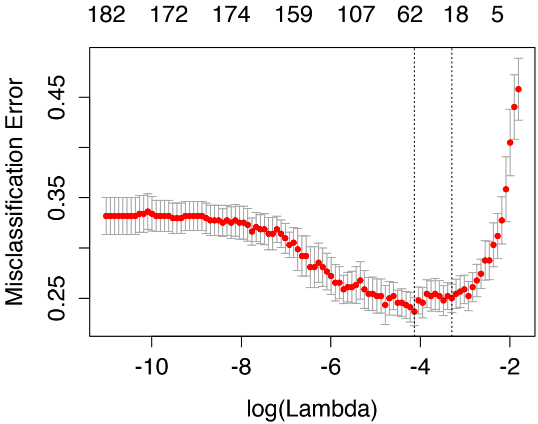
For sparse models, enough examples with respect to the number of relevant predictors

Least-squares regression



Lasso (logistic regression)

$n = 271$ examples



Data-analysis problems

Signal structure

Methods
- General techniques
- Denoising
- Signal recovery
- Signal separation
- Regression
- Compression / dimensionality reduction
- Clustering

When is the problem well posed?

Theoretical analysis

Optimization problem

$$\begin{aligned} \text{minimize} \quad & f(\tilde{x}) \\ \text{subject to} \quad & Ax = y \end{aligned}$$

f is a nondifferentiable convex function

Examples: ℓ_1 norm, nuclear norm

Aim: Show that the original signal x is the solution

Dual certificate

Subgradient of f at x of the form

$$q := A^T v$$

For any h such that $Ah = 0$

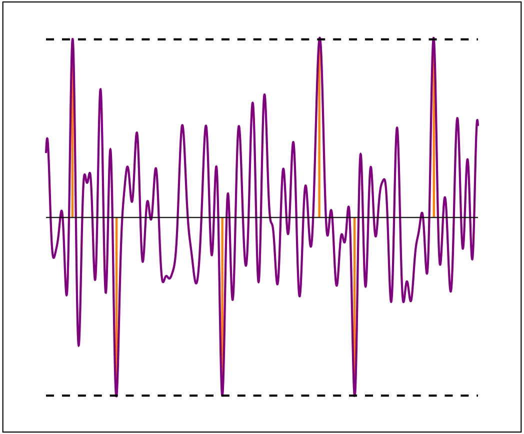
$$\langle q, h \rangle = \left\langle A^T v, h \right\rangle = \langle v, Ah \rangle = 0$$

$$f(x+h) \geq f(x) + \langle q, h \rangle = f(x)$$

Certificates

	Subgradient	*Row space of A*
Compressed sensing	$\text{sign}(x) + z$, $\|\|z\|\|_\infty < 1$	Random sinusoids
Super-resolution	$\text{sign}(x) + z$, $\|\|z\|\|_\infty < 1$	Low-pass sinusoids
Matrix completion	$UV^T + Z$, $\|\|Z\|\| < 1$	Observed entries

Certificate for compressed sensing



Certificate for super-resolution