

SPARSE SOLUTIONS
of LINEAR EQUATIONS
SPARSE^{*} MODELING
of SIGNALS & IMAGES

•

Alfred M. BRUCKSTEIN
David L. DONOHU
Michael ELAD

SIAM REVIEW, March 2009

OVERVIEW

A LINEAR SYSTEM

$$\underbrace{\begin{bmatrix} | & | & & | \\ a_1 & a_2 & \dots & a_M \\ | & | & & | \end{bmatrix}}_{A_{(N \times M)}} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix}$$

is underdetermined ($M > N$), and from the many possible \bar{x} 's we want to select one that best makes sense in view of prior information.

A MEASURE OF "DESIRABILITY" for \bar{x} is defined as $J_*(\bar{x})$ and we solve

Problem(*): $\min_{\bar{x}} J_*(\bar{x})$ subject to $\bar{r} = A\bar{x}$

$J_*(\bar{x})$ can be

0) # of nonzero entries in \bar{x} : $J_0(\bar{x})$

1) $J_1(\bar{x}) = \sum_{i=1}^M |x_i| = \ell_1\text{-norm of } \bar{x}$

2) $J_2(\bar{x}) = \bar{x}^T \bar{x} = \left(\sum_{i=1}^M |x_i|^2 \right)^{1/2} = \ell_2\text{-norm of } \bar{x}$

p) $J_p(\bar{x}) = \left(\sum_{i=1}^M |x_i|^p \right)^{1/p} = \ell_p\text{-norm of } \bar{x}$

The 2009-paper survey a set of wonderful results of the research community providing

EFFICIENT & STABLE ALGORITHMS to address SPARSITY-driven INVERSE PROBLEMS (with $J_0(\bar{x}), J_1(\bar{x})$) and their APPLICATIONS.

The message overview is:

- if A has some good properties ($\mu(A)$ -small) then if \bar{x} solving $A\bar{x} = \bar{v}$ exists with $J_0(\bar{x}) < \frac{1}{2}(1 + 1/\mu(A))$ then Problem (0) and Problem (1) have the same solution.
- both greedy algorithms and linear programming will solve Problem (0) in this case

Further messages:

- there are practical ways to model signals as sparse combinations of atom vectors $\{\mathbf{a}_i\}$ determined from training sets of signals (the K-SVD).
- many applications in signal/image denoising, compression, compressive sensing structure analysis are discussed.

I shall not repeat here the contents of the 2009-paper, but rather present my own path to sparsity and a connection to OVERPARAMETRIZED VARIATIONAL METHODS that I personally find very exciting & interesting

A ROAD-MAP:

- MUSIC & Resolution of Echos

RO Schmidt (1979)

Bruckstein Shan Kailath (1985)

- OVERPARAMETERIZED Variational methods

OpticFlow (2006) Nir, Bruckstein, Kimmel

Local Modeling (2007) Nir, Bruckstein

(2013) Shem-Tov, Rosman, Adir
Kimmel, Bruckstein

- MOVING LEAST SQUARES

Savitzky Golay (1964)

Lancaster, Salkauskas (1981)

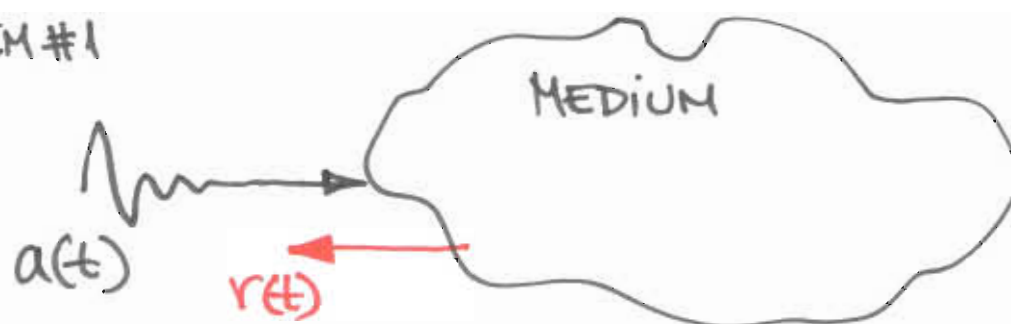
- NONLOCAL OVERPARAMETRIZED VARIATIONAL
method vs SPARSITY Based Solutions

Giryas, Elad, Bruckstein

(2014-2020??)

MUSIC & ECHO RESOLUTION

PROBLEM #1



$$r(t) = \sum_{i=1}^K s_i a(t - \theta_i) + n(t) \quad \text{(noise)}$$

Sampling $r(t)$ at N points in time we have

$$\bar{r} = \begin{bmatrix} r(t_1) \\ r(t_2) \\ \vdots \\ r(t_N) \end{bmatrix} = \begin{bmatrix} | & | & & | \\ a(\theta_1) & a(\theta_2) & \dots & a(\theta_K) \\ | & | & & | \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_K \end{bmatrix} + \begin{bmatrix} n(t_1) \\ n(t_2) \\ \vdots \\ n(t_N) \end{bmatrix}$$

with $\bar{a}(\theta) \triangleq [a(t_1 - \theta) \ a(t_2 - \theta) \ \dots \ a(t_N - \theta)]$
but this can be viewed as

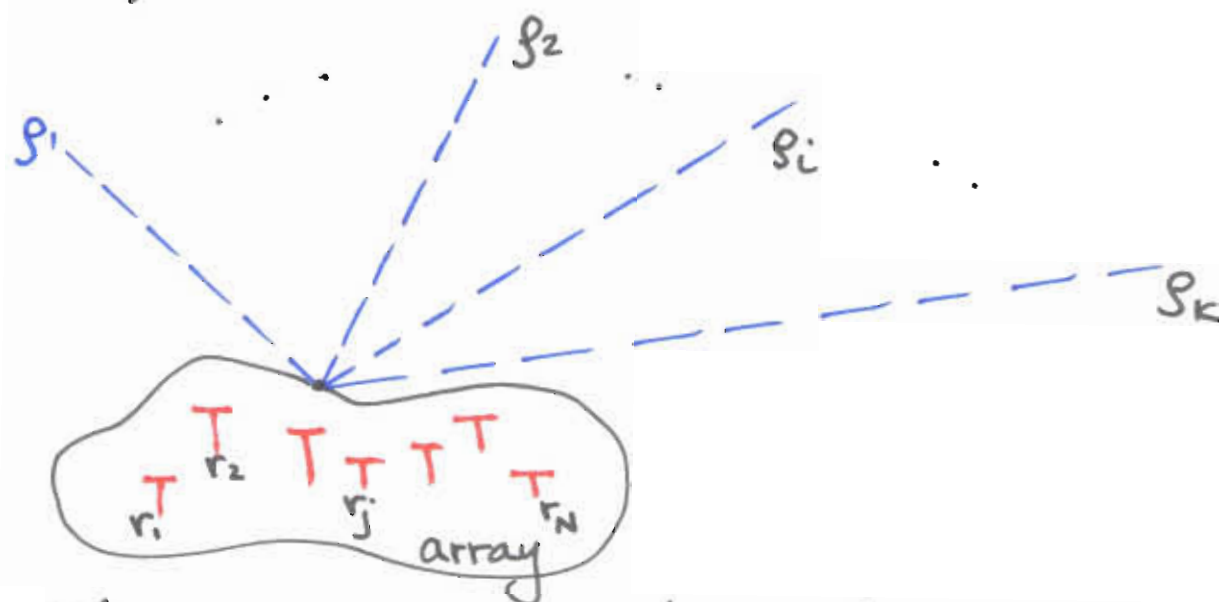
$$\bar{r} = \underbrace{\begin{bmatrix} \dots & \dots & a(\theta) & \dots & \dots \end{bmatrix}}_{\text{all } \theta\text{'s possible (finely discretized to } M \text{ values)}} \begin{bmatrix} \vdots \\ s_1 \\ \vdots \\ s_2 \\ \vdots \\ \vdots \\ s_K \\ \vdots \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_N \end{bmatrix}$$

a sparse vector with only K nonzero entries

Given \bar{r} estimate $\{\theta_1, \theta_2, \dots, \theta_K\}$.

PROBLEM #2

We consider an antenna array recording signals from radiation sources located at K directions



The signals recorded by the N antennas are given by

$$\bar{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix} = \begin{bmatrix} a(\theta_1) & a(\theta_2) & \dots & a(\theta_K) \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_K \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_N \end{bmatrix}$$

where:

- p_i is the (random) signal from radiation source i
- $a(\theta)$ - the array response to a "calibration" signal in direction θ

$\{a(\theta)\}_{\theta \in [0, 2\pi]}$ is the "array manifold", a "curve" in N -dimensional space.

From \bar{r} , here too, we want to recover $\{\theta_1, \theta_2, \dots, \theta_K\}$.

For both problems discussed above the recorded response is modeled as a superposition of K vectors selected from a θ -parametrized set of vectors $\{a(\theta)\}$. The θ set can be "finely sampled" to M discrete values and then the set $\{a(\theta)\}$ can be displayed as a matrix $A_{(N \times M)}$ and we have

$$\vec{r} = \underbrace{\begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & a(\theta_i) & \dots & \dots & \dots \end{bmatrix}}_{A_{N \times M}} \begin{bmatrix} \bullet \\ \bullet \\ \bullet \\ \bullet \\ \bullet \end{bmatrix} + \bar{m}$$

a long vector with only K nonzero entries.

The matrix A obeys (usually) from physical constraints or by design, that any N columns of it form a full rank square matrix, i.e. no column can be replaced by a linear combination of fewer than N others.

R.O. Schmidt's genial MUSIC algorithm (1979)

assumes we are recording repeated randomly drawn (in terms of the noise \underline{n} and the \underline{p} 's) response vectors $\{\bar{r}^t\}_{t=1,2,3 \dots 1000000}$, for a fixed set of θ 's: $\{\theta_1, \theta_2 \dots \theta_K\}$.

Then we can do:

- Estimate the autocorrelation matrix

$$\begin{aligned} \mathcal{R}_r &= E \bar{r} \bar{r}^T = A(\theta_1, \theta_2 \dots \theta_K) E \underline{p} \underline{p}^T A^T(\theta_1, \theta_2 \dots \theta_K) + E \underline{n} \underline{n}^T \\ &= A \cdot \mathcal{R}_p A^T + \underbrace{\sigma_n^2 \mathbf{I}}_{\mathcal{R}_n} \end{aligned}$$

- From this the eigenvalues of \mathcal{R}_r will be:
 $N-K$ eigenvalues $= \sigma_n^2$ and K higher ones since \mathcal{R}_p is assumed to be a full rank matrix (i.e. the sources of echos or radiation are uncorrelated or at least not fully correlated)
from this we estimate σ_n^2 and compute $\mathcal{R}_r - \sigma_n^2 \mathbf{I}$

- Now $\mathcal{R}_r - \sigma_n^2 \mathbf{I}$ has $N-K$ zero eigenvalues with a nullspace spanned by $N-K$ orthonormal vectors $\{\bar{v}_{k+1}, \bar{v}_{k+2} \dots \bar{v}_N\}$.

- Hence we know that

$$(\mathbb{R}_r - \sigma_r \mathbb{I}) \underline{v}_j = \underline{0} \quad \text{for } j = k+1, k+2, \dots, N$$

$$\text{or } A \mathbb{R}_p A^T \underline{v}_j = 0$$

$$\Rightarrow \boxed{A^T \underline{v}_j = \underline{0}} \quad \text{for } j = k+1, k+2, \dots, N$$

Therefore $\langle a(\theta_i), \underline{v}_j \rangle = 0$ for $i = 1, 2, \dots, K$
 $j = k+1, k+2, \dots, N$

- Now search for all θ 's for which

$$\langle a(\theta), \underline{v}_j \rangle = 0 \quad (j = k+1, k+2, \dots, N)$$

by plotting for example the function

$$\Psi(\theta) = \frac{1}{\sum_{j=k+1}^N \langle a(\theta), \underline{v}_j \rangle^2}$$

and select the K places where $\Psi(\theta)$ peaks
 (theoretically it should become ∞ !).

Wonderful, isn't it?



But, what happens if we have only one (or very few!) response \underline{r} and we cannot estimate \mathbb{R}_r at all!

In the echos example radar engineers are taught to use the "optimal" matched filter to the signal $a(t)$. This filter correlates $r(t)$ with $a(t-\theta)$ for all θ 's in the range of interest and the response

$$\Psi(\theta) = \int_{\Omega_t} r(t) a(t-\theta) dt \quad (*)$$

provides estimates for θ_i 's as peaks of $\Psi(\theta)$.

(*) radar engineers are very happy with this because $\Psi(\theta)$ is the result of a convolution operator, readily implementable as a fixed time-invariant filter!

We can take this idea for the general case and compute for all $a(\theta)$ the inner products $\langle \bar{r}, a(\theta) \rangle = \Psi_{\bar{r}}(\theta)$

Therefore in the one-shot case we do

- for all $\{\theta\}$, compute
 $\langle \bar{r}, a(\theta) \rangle = \psi_r(\theta)$
select the K maximal values of $\psi_r(\theta)$
as the estimates for $\theta_1, \theta_2 \dots \theta_K$.

(The Thresholding Algorithm)

or we can also proceed as follows

1. for all $\{\theta\}$ compute
 $\langle \bar{r}, a(\theta) \rangle = \psi_r(\theta)$
select the maximum value of $\psi_r(\theta)$
~ it provides θ_1 , then do

$$\bar{r} - \langle \bar{r}, a(\theta_1) \rangle a(\theta_1) \rightarrow \bar{r}^{\text{next}}$$

2. Now for all $\{\theta\} \neq \theta_1$ compute
 $\langle \bar{r}^{\text{next}}, a(\theta) \rangle = \psi_{r^{\text{next}}}(\theta)$

select maximum value

~ it provides θ_2 then do

3.

as before

... ..

OK

etc.

This is the

MATCHING PURSUIT ALGORITHM

For both the echo resolution and direction of arrival estimation problems we were led to the need to solve the problem

$$\begin{array}{l} \text{Find } \bar{x} \text{ so that } J_0(\bar{x}) \leq K \\ \text{subject to } A\bar{x} \approx r \\ \text{or } J_2(\bar{r} - A\bar{x}) \approx \sigma_{\text{Noise}} \end{array}$$

← sparsity constraint

As we know that such problems are variations on the basic sparse recovery problems of the type discussed in the 2009 paper.

However if we would prefer the variational approach we could also write

$$\min \| \bar{x} \|_1 + \lambda \| \bar{r} - A\bar{x} \|_2$$

or in the continuous case

$$\min_{x(\theta)} \Psi\{x(\theta)\} = \int_{\Omega_\theta} |x(\theta)| d\theta + \lambda \left(\int_{\Omega_t} \left(r(t) - \int_{\Omega_\theta} a(t, \theta) x(\theta) d\theta \right)^2 dt \right)$$

This continuous case variational problem reminds us of the wealth of activity in variational denoising using total variation constraints as the smoothness terms.

Indeed in the classical ROF framework we are given a noisy signal $r(t)$
 $r(t) = m(t) + n(t)$:

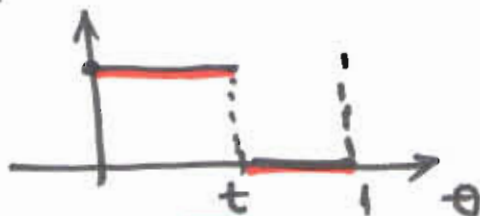
and we want to recover $m(t)$ we do:

$$\min_m \Psi\{m\} = \int_{\Omega_t} (r(t) - m(t))^2 dt + \lambda \int_{\Omega} |m'(t)| dt$$

$$= \int_{\Omega_t} (r(t) - \int_0^t m'(\theta) d\theta)^2 dt + \lambda \int_{\Omega} |m'(t)| dt =$$

$$= \int_{\Omega_t} (r(t) - \int_{\Omega_t} a(t, \theta) m'(\theta) d\theta)^2 + \lambda \int_{\Omega} |m'(t)| dt$$

where $a(t, \theta)$:



Rewritten in terms of $x(\theta) \triangleq \frac{d}{d\theta} m(\theta)$ over the range $\theta \in \Omega_\theta (= [0, 1]$ say) we have that the recovery of $m(t)$ can proceed as follows

$$\min_{x(\theta)} \Psi\{x(\theta)\} = \int_{\Omega_t} \left(r(t) - \int_{\Omega_\theta} a(t, \theta) x(\theta) d\theta \right)^2 dt + \lambda \int_{\Omega_\theta} |x(\theta)| d\theta$$

and for the optimal $x(\theta)$ we simply integrate to obtain $m(t)$.

In the discrete case what we have done is to define

$$\vec{r} = \underbrace{\begin{bmatrix} \vdots & \vdots & \vdots & \vdots \end{bmatrix}}_{\leftarrow A \rightarrow} \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix} \times \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix} + \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}$$

This circle of ideas leads to a sparsity based approach to variational methods aimed to recover OVERPARAMETRIZED signals or images.

A signal is overparametrized if it is modeled as follows:

$$m(t) = \sum_{i=1}^K \underbrace{x_i(t)} \cdot \underbrace{\varphi_i(t)}$$

where $\varphi_i(t)$ are given functions
(from modeling the real life problem)

and $x_i(t)$ are "coefficient" functions
usually known to be piecewise constant
over some partition of the domain of t .

If we observe $m(t)$ with additive noise

i.e.
$$r(t) = m(t) + n(t)$$

we want to recover $m(t)$ (or in fact $x_i(t)$, $i=1,2,\dots,K$) from $r(t)$.

The variational approach which is "natural" in this case is: define the functional

$$\Psi\{x_i(t) | i=1,2,\dots,k\} = \int_{\Omega_t} \left(r(t) - \sum_{i=1}^k x_i(t) \phi_i(t) \right)^2 dt + \left\{ \begin{array}{l} \sum_{i=1}^k \int_{\Omega_t} |x_i'(t)| dt \\ \text{or any other} \\ \text{penalty} \\ \text{encouraging} \\ \text{piecewise constant} \\ \text{functions } x_i(t) \end{array} \right.$$

Using this approach we had good results of denoising, but we did not obtain good estimates for the $x_i(t)$ -functions. Paradoxically the "Euler-Lagrange" did a good job in getting the sums $\sum x_i(t) \phi_i(t)$ but not in recovering piecewise constant $x_i(t)$'s (which would have provided us good SEGMENTATION RESULTS).

This was due to the lack of good penalties HERE.

The NEW APPROACH:

Consider a simple, discrete time example of a signal which is piecewise linear.

$$m(t) = 1 \cdot x_1(t) + t \cdot x_2(t)$$

$$\text{or } m(i) = 1 \cdot x_1(i) + i \cdot x_2(i)$$

In this case we shall write the functional

$$\Psi\{x_1, x_2\} = \left\| \bar{r} - \begin{bmatrix} 1 & \dots & 1 \end{bmatrix} x_1 - \begin{bmatrix} 1^2 & 3 \dots N \end{bmatrix} x_2 \right\|_2 +$$

$+ \lambda \left\{ \begin{array}{l} \text{a penalty term making the} \\ \text{vectors } x_1 \text{ and } x_2 \text{ piecewise} \\ \text{constant \& changing in unison!} \end{array} \right\}$

$$= \left\| \bar{r} - \begin{bmatrix} 1 & \dots & 1 \\ 1^2 & 3 \dots N \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\|_2 +$$

$+ \lambda \left[\begin{array}{l} \text{Functional Making the Vectors} \\ \Delta x_1 \text{ and } \Delta x_2 \\ \text{Sparse with identical sparsity} \\ \text{patterns} \end{array} \right]$

The vectors Δx_1 and Δx_2 are given by

$$\begin{bmatrix} 1 & & & & 0 \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ 0 & & & & 1 \end{bmatrix} \begin{bmatrix} x_1^{1/2} \\ x_2^{1/2} \\ \vdots \\ x_N^{1/2} \end{bmatrix} = \Delta x_{1/2}$$

$$\text{Hence } \begin{bmatrix} x_1^{1/2} \\ \vdots \\ x_N^{1/2} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & & & & 0 \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}}^H \Delta x_{1/2}$$

Now we can write the \dagger optimization problem to be solved as follows

$$\text{minimize } \left\| r - \underbrace{\begin{bmatrix} I & H \\ & [I^2 \dots I^N] H \end{bmatrix}}_A \begin{bmatrix} D^1 \\ D^2 \end{bmatrix} \right\|_2 + \lambda \left\| D^1 \cdot D^1 + D^2 \cdot D^2 \right\|_0$$

where $D^1 = \Delta x_1$, $D^2 = \Delta x_2$

$$\text{and } D \cdot D = \begin{bmatrix} d_1^2 \\ d_2^2 \\ \vdots \\ d_N^2 \end{bmatrix}$$

Term enforcing joint sparsity pattern

In this form the optimization is a structured-sparsity problem of an almost standard type.

We (Raja Giryes, Michael Elad & B) tested several cases of 1D piece-wise linear and 2D-image piecewise planar fitting/denoising problems using this approach and the results are very good, especially in enforcing excellent segmentations (as opposed to the Euler-Lagrange based gradient descent methods!).

We shall report on this in a detailed paper in the near future.

We next show some examples

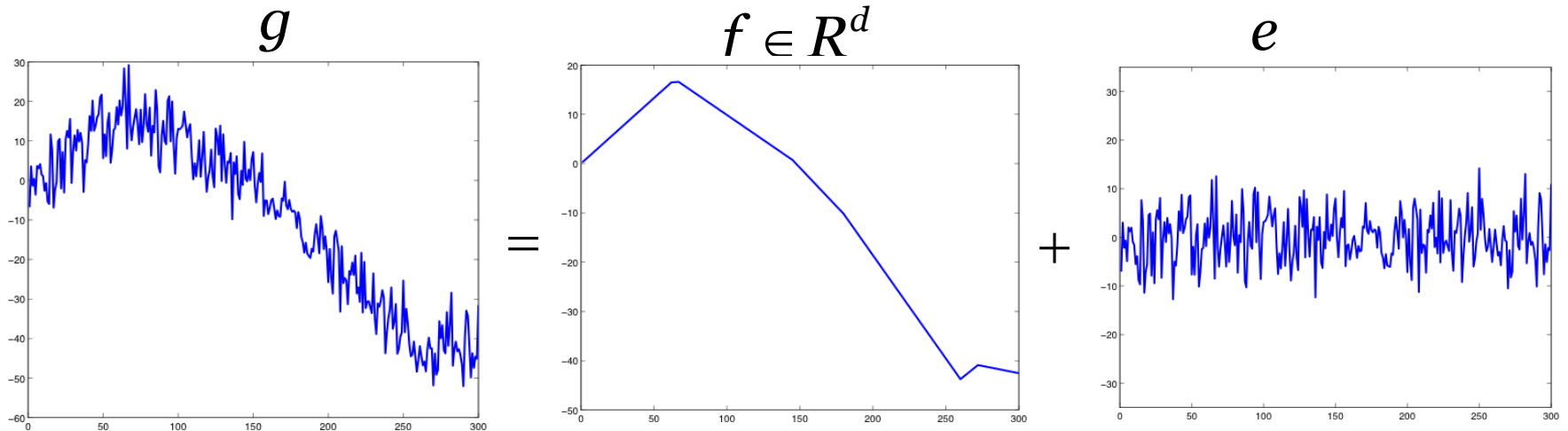
1) Examples for 1D signal
denoising & segmentation

2) Examples of Image
approximations by piecewise
planar patches

showing DENOISING
SEGMENTATION
INPAINTING

The Line Fitting Problem

The Problem:



The Model: overparametrized representation for piecewise linear function


$$f = a + Xb = [I \quad X] \begin{bmatrix} a \\ b \end{bmatrix}, \quad X = \text{diag}(1, \dots, d)$$

Sparsity Based Solution

Notice that $\Omega_{\text{DIF}}a$ and $\Omega_{\text{DIF}}b$ should be sparse at the same location.

\Rightarrow Solve:

$$\min_{a,b} \left\| g - \begin{bmatrix} \text{I} & \text{X} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \right\|_2^2 \quad \text{s.t.} \quad \left\| |\Omega_{\text{DIF}}a| + |\Omega_{\text{DIF}}b| \right\|_0 \leq k,$$

Structured
sparsity 

where k is the number of jumps.

If k is unknown but the noise energy $\|e\|_2^2$ is known, solve

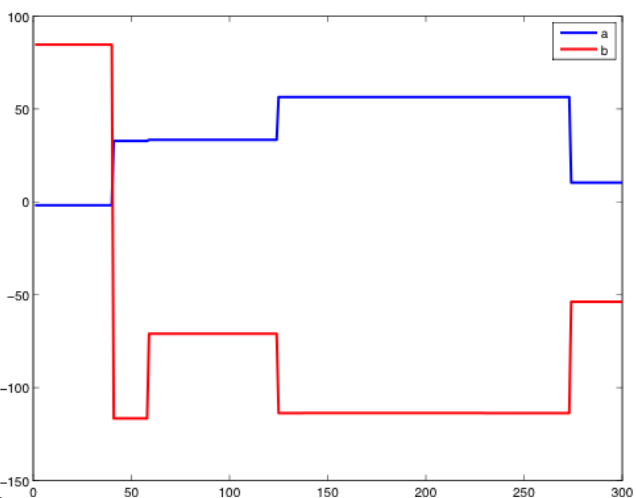
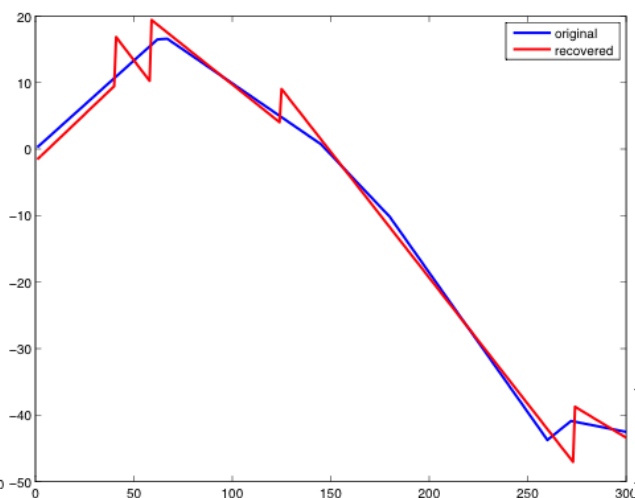
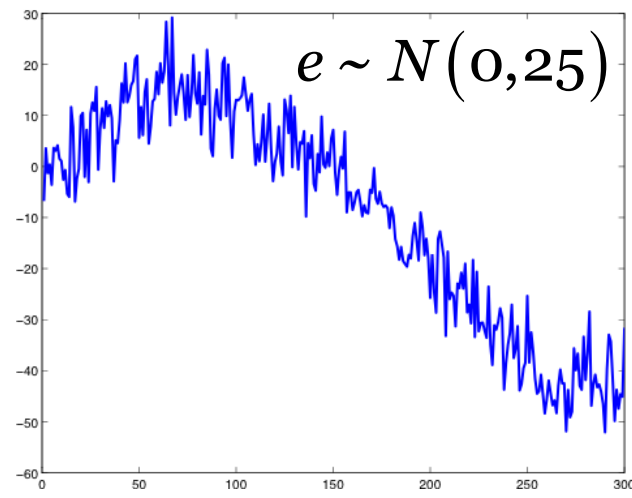
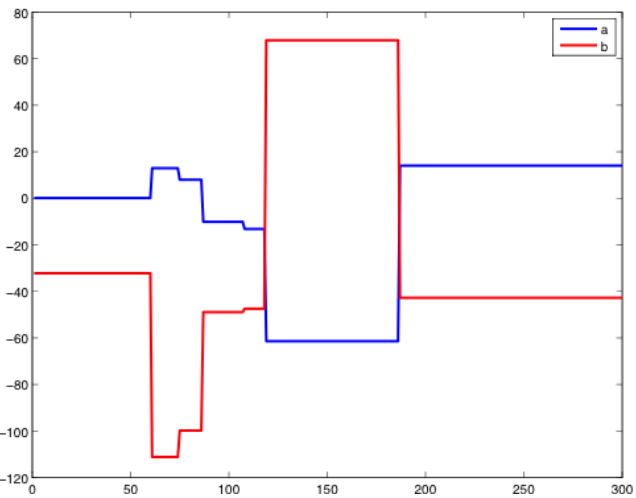
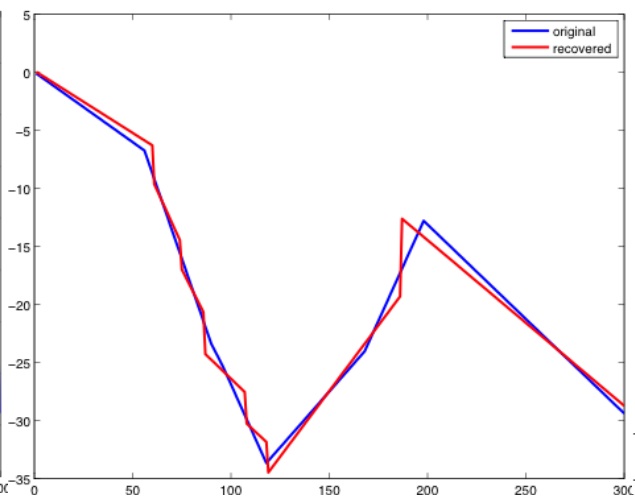
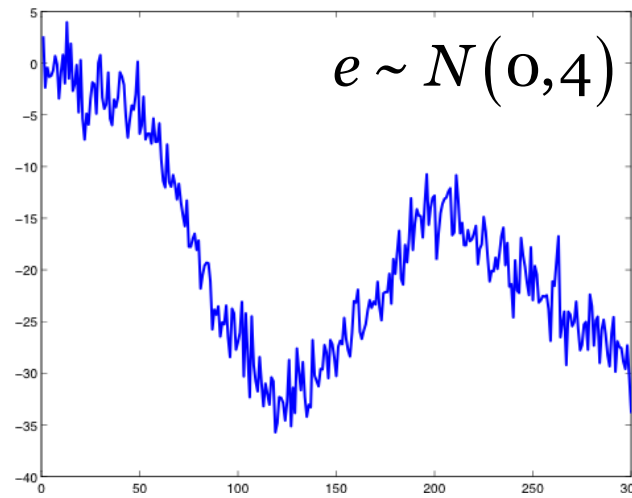
$$\min_{a,b} \left\| |\Omega_{\text{DIF}}a| + |\Omega_{\text{DIF}}b| \right\|_0 \quad \text{s.t.} \quad \left\| g - \begin{bmatrix} \text{I} & \text{X} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \right\|_2^2 \leq \|e\|_2^2.$$

The above problems belongs to the analysis (co)sparse framework.

We use the GAPN algorithm [\[Nam, Davies, Elad and Gribonval, 2013\]](#)

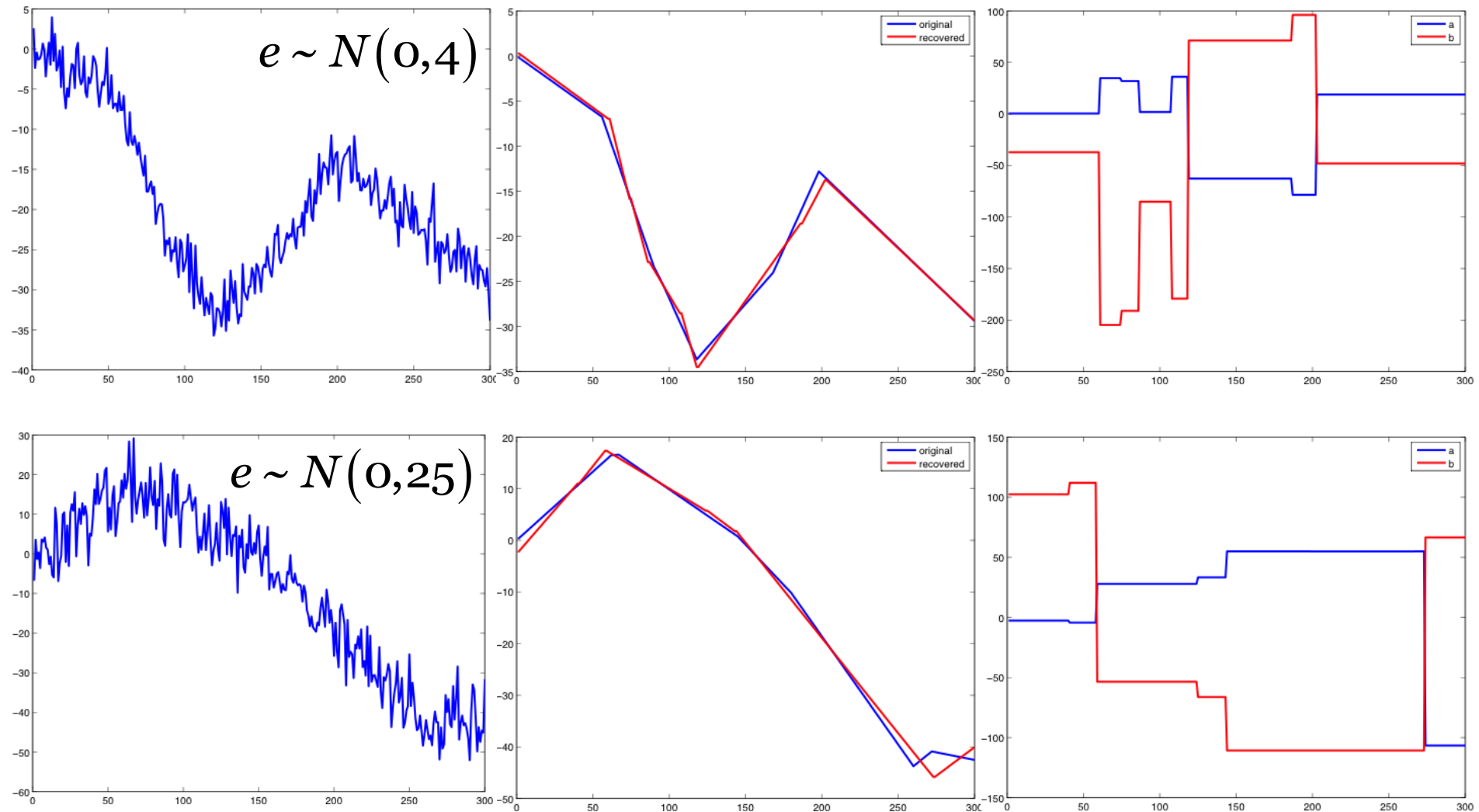
We modify it to support structured sparsity.

Line Fitting Experiment



Solving the Jumps Problem

Add a constraint for continuity between jumps to the minimization problem



From 1D to 2D

We set

$$f = a + Xb_1 + Yb_2 = \begin{bmatrix} I & X & Y \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix},$$

There are many options to extend Ω_{DIF} to 2D.

We consider two:

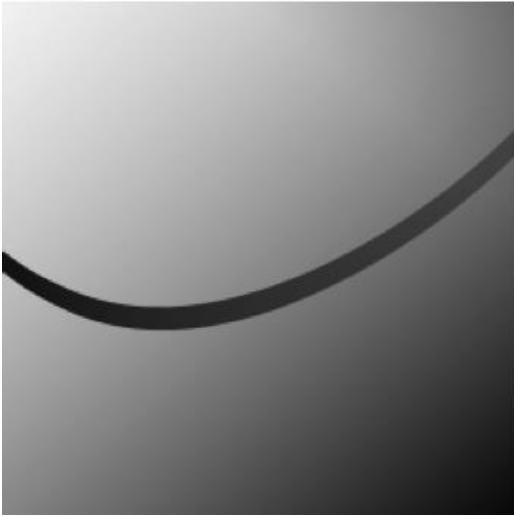
1) Standard: horizontal and vertical derivatives

$$\begin{bmatrix} 1 & -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

2) Standard + diagonal derivatives

$$\begin{bmatrix} 1 & -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Denoising Example



Original



Ω_{DIF} – Standard, PSNR=39.09



Noisy, $\sigma=20$

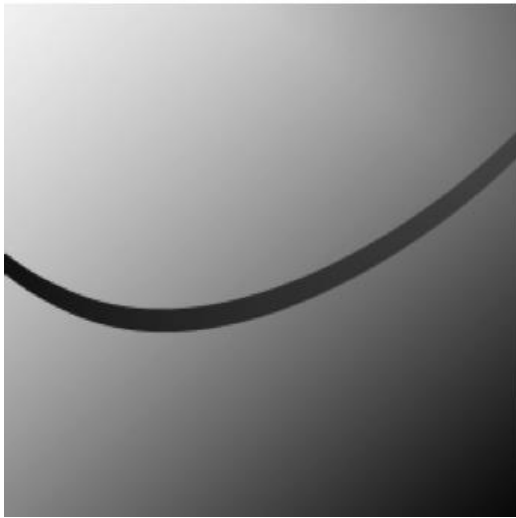


Ω_{DIF} – With Diagonal, PSNR=39.57

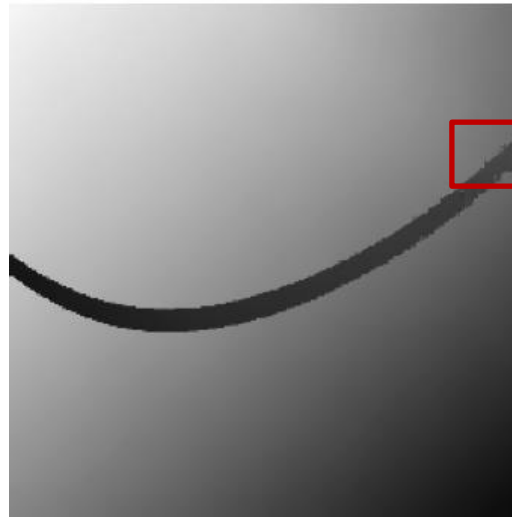


ROF, PSNR=33.21
[Rudin, Osher and Fatemi, 1992]

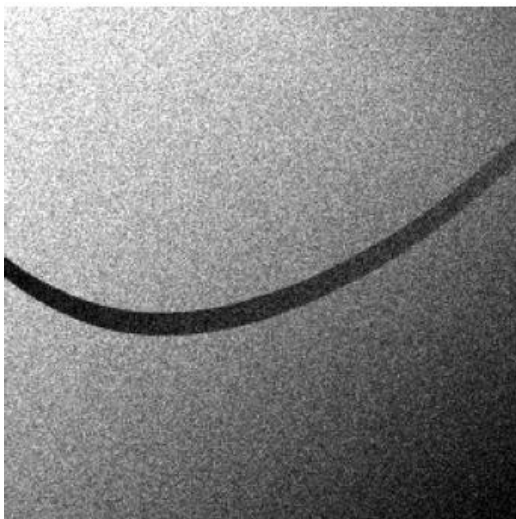
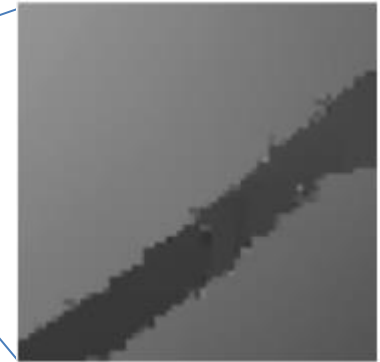
Denoising Example



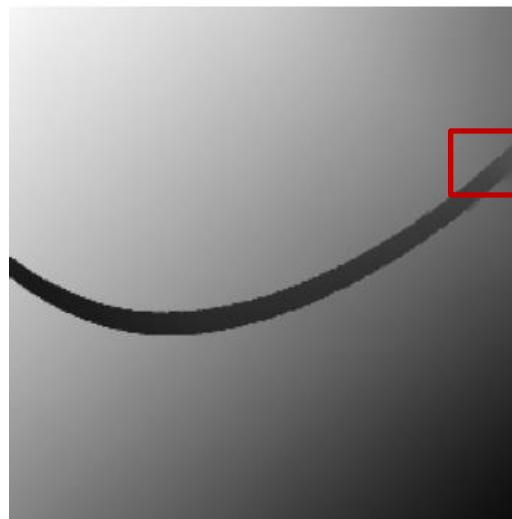
Original



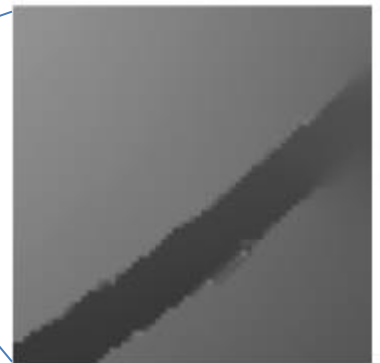
Ω_{DIF} – Standard



Noisy, $\sigma=20$



Ω_{DIF} – With Diagonal



Denoising Example



Original



Ω_{DIF} - Standard, SNR=29.6



Noisy, $\sigma=20$



ROF, PSNR=30.28

ROF gets better SNR since the new method assumes a piecewise linear image and therefore smoothes regions in the image.

Is this useful?

Segmentation



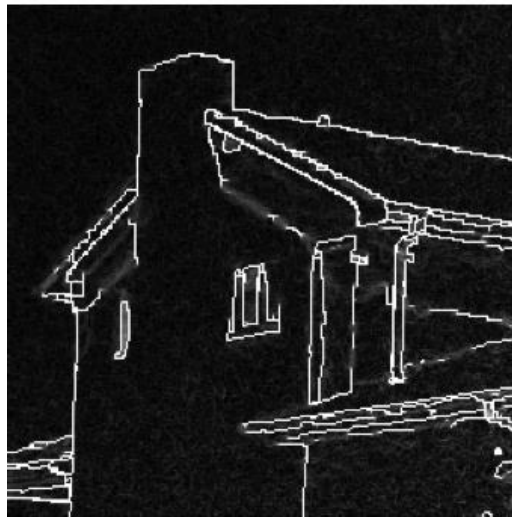
Original



Gradient Map of
the Original Image

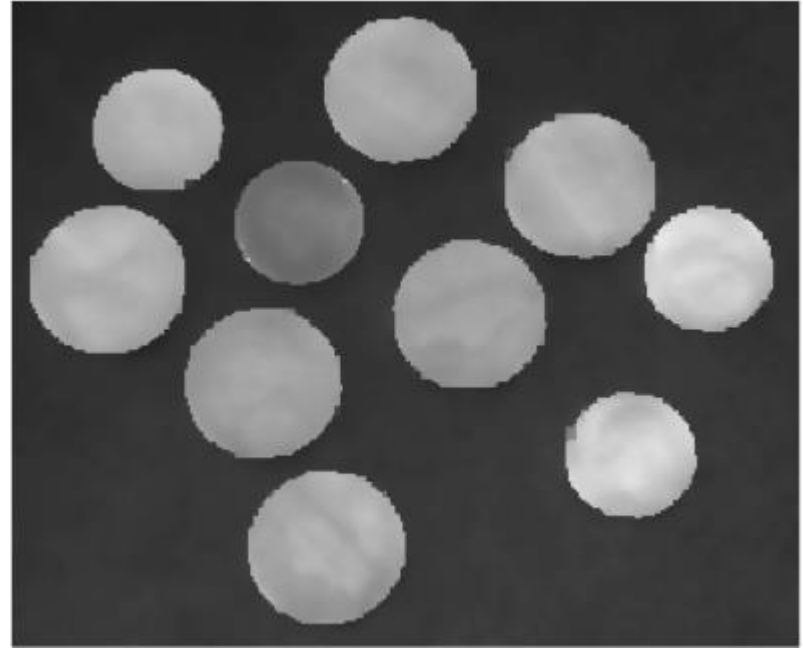
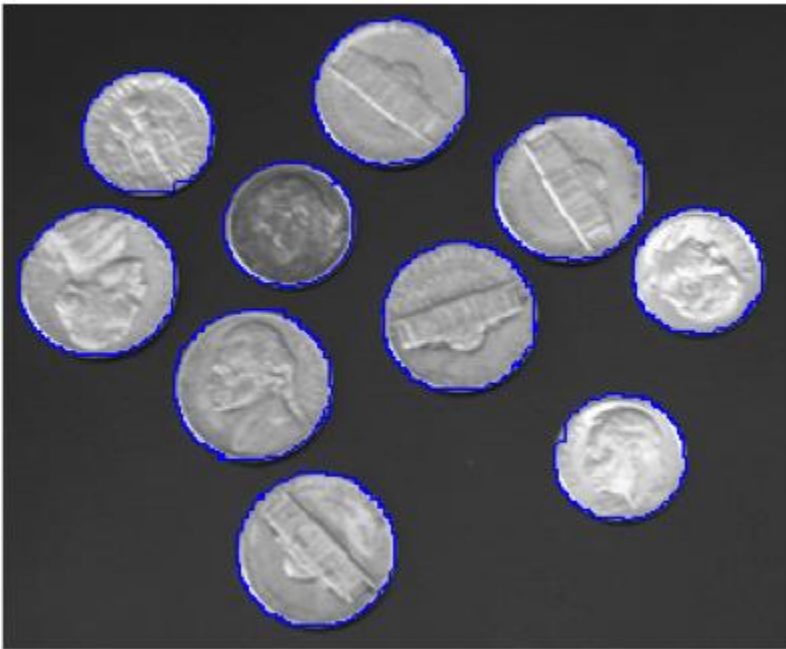


Ω_{DIF} – Standard Recovery

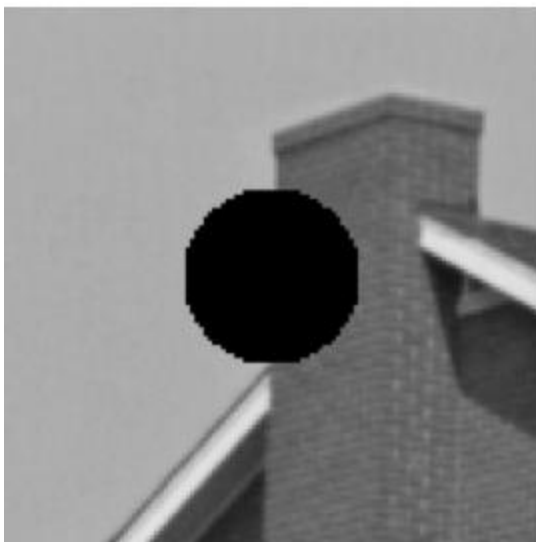


Gradient Map of the Piecewise
Constant Coefficients of the
Recovered Image

Segmentation



Geometrical Inpainting



Geometrical Inpainting



With standard
derivatives



With diagonal
derivatives

