Variable selection and outlier detection as a Mixed Integer Progam (MIP) and Image segmentation

Stéphane Canu
asi.insa-rouen.fr/enseignants/~scanu

S4D 2018 : Research Summer School on Statistics for Data Science

Caen, June 18, 2018

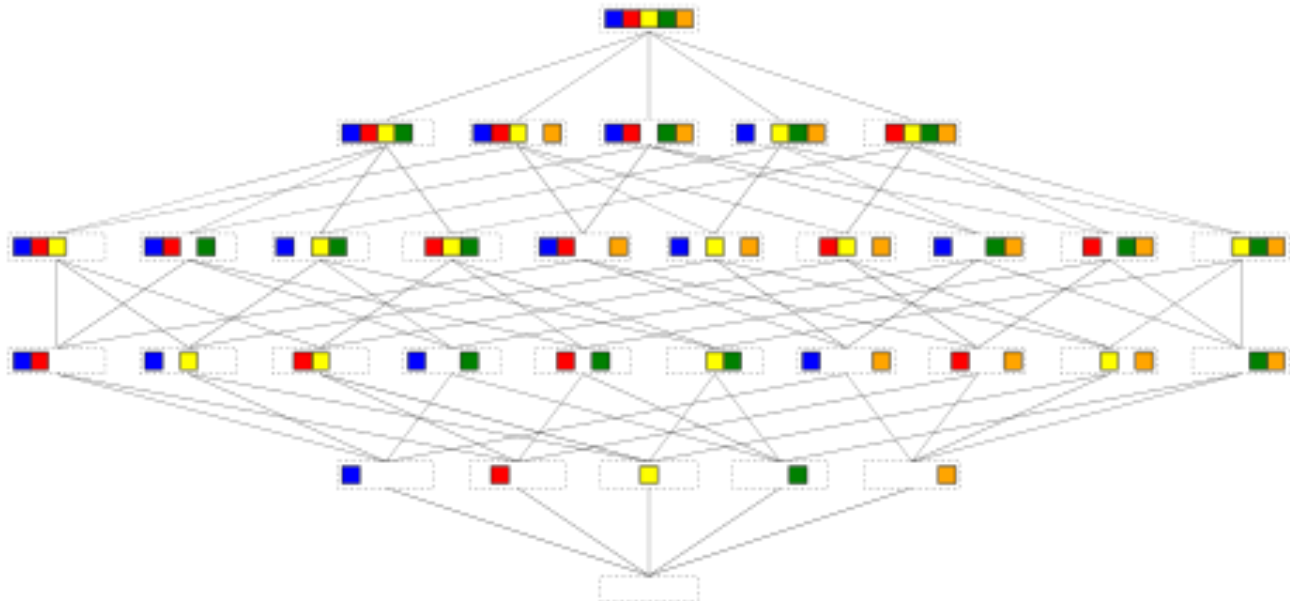
Road map

1. Examples of combinatorial problems in machine learning
2. Mixed integer (binary) programming (MIP)
3. L_0 proximal algorithm
4. Implementation
5. MIP for image processing

NP Hard =

Variable selection

$$f(x_1, \ldots, x_j, \ldots, x_p) = \sum_{j=1}^{p=5} x_j w_j$$

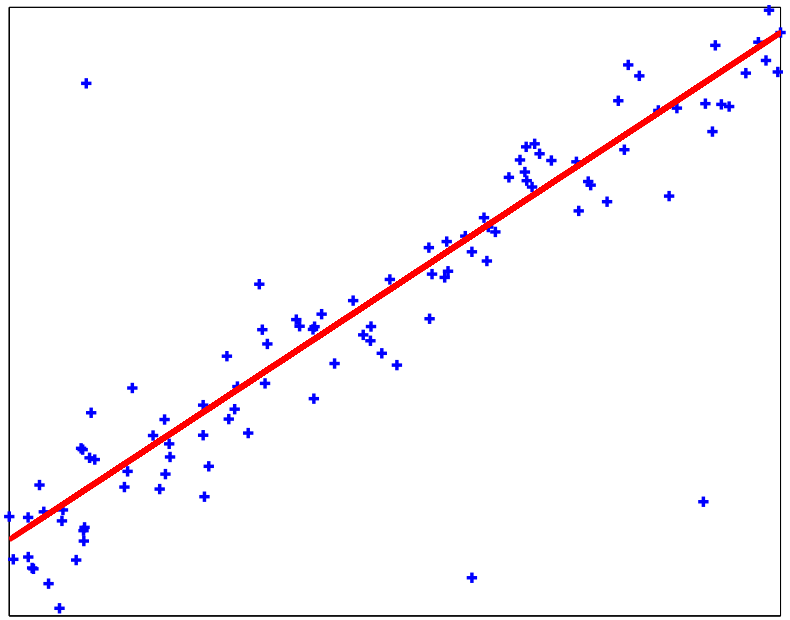


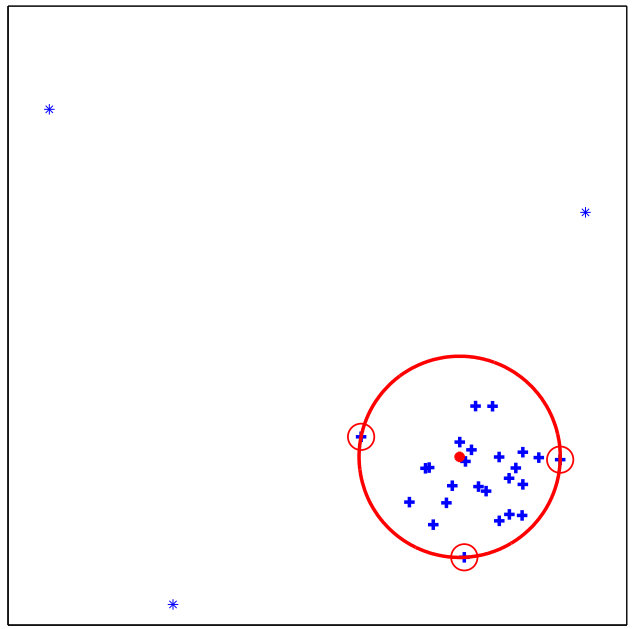
e.g. gene selection from microarray data, find w sparse such that $X\mathrm{w} = \mathrm{y}$

Fit the data and remove useless variables

Enumerate of all possible combinations and choose

Outlier detection





Fit the data and remove useless observations (outliers)

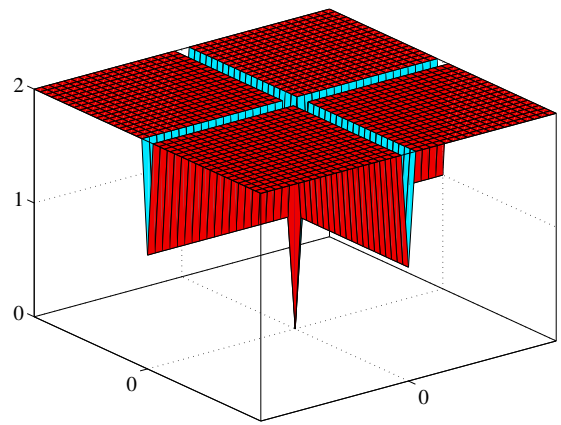
Enumerate of all possible point configurations and choose

What's common?

Example (The counting function)

$$\begin{array}{rccl} c: & \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & \mathrm{w} & \longmapsto & c(\mathrm{w}) = \text{ the number of nonzero components } w_i \text{ of } \mathrm{w} \end{array}$$

It is often called the 0-norm denoted by $c(\mathrm{w}) = \|\mathrm{w}\|_0$.



Minimize a nonconvex nonsmooth target function or constraint

Nonconvex Nonsmooth problems in machine learning

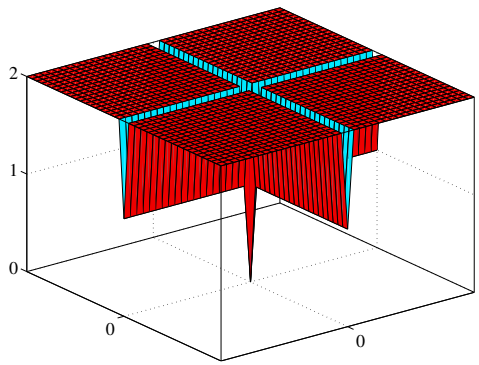
Many lattice based problems

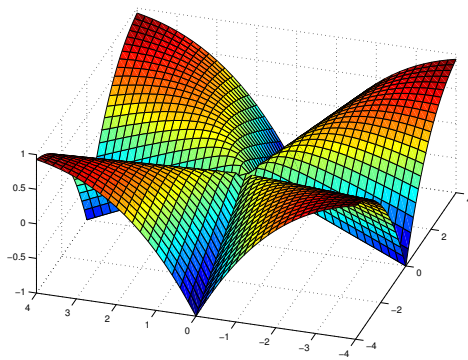
- variable selection, outlier detection, clustering,
- image processing, total variations,
- discrete artificial vision,
- sensor placement,
- distribution factorization,
- low rank factorization.

3 ways of solving combinatorial problems

- local optimization (in general)
  - Continuous relaxations (L_1 penalty, DC...)
  - Combinatorial algorithms (greedy search, spanning tree...)
- global optimization
  - Mixed integer programming (difficult to scale to large problems)

Road map





Variable selection with binary variables

Definition (the least square variable selection problem)

$$\begin{cases} \min\limits_{w \in \mathbb{R}^p} & \|Xw - y\|^2 & \longleftarrow \text{fit the data} \\ \text{s.t.} & \|w\|_0 \leq k & \longleftarrow \text{with k variables} \end{cases}$$

- introduce p new binary variable $z \in \{0,1\}^p$
- for useless variables: $z_j = 0 \Leftrightarrow w_j = 0$ $\rightarrow$ a coupling mechanism
- $\|w\|_0 = \sum_{j=1}^{p} z_j$

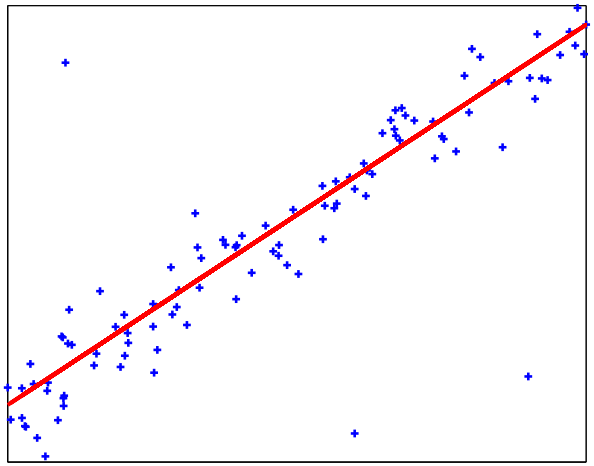
Definition (the LS variable selection problem with binary variables)

$$\begin{cases} \min\limits_{w \in \mathbb{R}^p, z \in \{0,1\}^p} & \|Xw - y\|^2 \\ \text{s.t.} & \|w\|_0 = \sum_{i=1}^{p} z_i \leq k \\ & z_j = 0 \Leftrightarrow w_j = 0, \quad j = 1, p \end{cases}$$

Outlier detection with binary variables

Introducing outliers variables $\mathrm{o} \in \mathbb{R}^n$

$$\mathrm{y} = X\mathrm{w} + \varepsilon + \mathrm{o}, \qquad \mathrm{o}_i = \begin{cases} y_i - \mathrm{x}_i^t \mathrm{w} & \text{if } i \text{ outlier} \\ 0 & \text{else} \end{cases}$$



The least square (trimmed) regression problem with k outliers [?]

$$\begin{cases} \min\limits_{\mathrm{w} \in \mathbb{R}^p, \mathrm{o} \in \mathbb{R}^n} & \frac{1}{2} \|X\mathrm{w} + \mathrm{o} - \mathrm{y}\|^2 \\ \text{s.t.} & \|\mathrm{o}\|_0 \leq k \end{cases}$$

Introduce binary variables

$$i = 1, n \qquad \begin{cases} t_i = 0 & (\mathrm{x}_i, y_i) \text{ is an outlier} & o_i \neq 0 \\ t_i = 1 & (\mathrm{x}_i, y_i) \text{ is NOT an outlier} & o_i = 0 \end{cases}$$

$$\|\mathrm{o}\|_0 = \sum_{i=1}^{n} (1 - t_i)$$

Bi robust regression

Variable selection AND outlier detection

$$\begin{cases} \min\limits_{\mathrm{w}\in\mathbb{R}^p} & \frac{1}{2}\,\|X\mathrm{w}-\mathrm{y}\|^2 \\ \text{s.t.} & \|w\|_0 \leq k_v \end{cases} \quad \text{and} \quad \begin{cases} \min\limits_{\mathrm{w}\in\mathbb{R}^p,\mathrm{o}\in\mathbb{R}^n} & \frac{1}{2}\,\|X\mathrm{w}+\mathrm{o}-\mathrm{y}\|^2 \\ \text{s.t.} & \|\mathrm{o}\|_0 \leq k_o \end{cases}$$

LS regression with variable selection AND outlier detection

Given k_v the number of variable required and k_o the number of outliers

$$\begin{cases} \min\limits_{w\in\mathbb{R}^p,\mathrm{o}\in\mathbb{R}^n} & \|Xw-y-\mathrm{o}\|^2 \\ \text{s.t.} & \|w\|_0 \leq k_v \\ & \|\mathrm{o}\|_0 \leq k_o. \end{cases} \tag{1}$$

Bi robust regression with binary variables

- Variables
- Outliers

$$\begin{cases} \min\limits_{w,o} & \|Xw - y - o\|^2 \\ \text{s.t.} & \|w\|_0 \le k_v \\ & \|o\|_0 \le k_o. \end{cases}$$

p binary variables $z_j \in \{0,1\}$

$$\|w\|_0 = \sum_{j=1}^{p} z_j \quad \text{and} \quad z_j = 0 \Leftrightarrow w_j = 0,$$

n binary variables $t_i \in \{0,1\}$

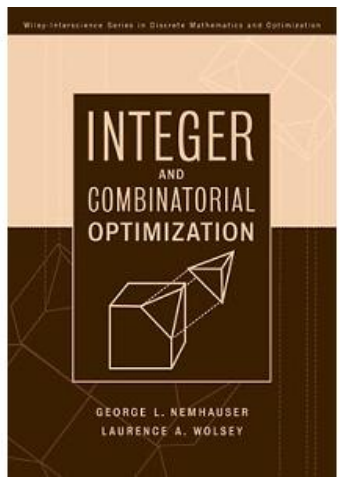
$$\|o\|_0 = \sum_{i=1}^{n} (1 - t_i) \text{ and } 1 - t_i = 0 \Leftrightarrow o_i = 0,$$

Bi robust regression

$$\begin{cases} \min\limits_{w \in \mathbb{R}^p, o \in \mathbb{R}^n, z \in \{0,1\}^p t \in \{0,1\}^n} & \|Xw - y - o\|^2 \\ \text{s.t.} & \|w\|_0 = \sum z_j \le k_v \\ & z_j = 0 \Leftrightarrow w_j = 0, \qquad j = 1, p \\ & \|o\|_0 = \sum t_i \le k_o \\ & 1 - t_i = 0 \Leftrightarrow o_i = 0, \qquad i = 1, n \end{cases} \tag{2}$$

So far...

- combinatorial problems can be formulated using binary variables
- we have mixed binary optimization problem
- How to solve them?
  - reformulations
  - towards stronger relaxations
  - nice initialization



Mixed integer linear program (MILP)

- linear cost
- linear constraints
- integer and continuous variables

Definition (mixed integer linear program – MILP (canonical form))

$$\begin{cases} \min\limits_{\mathrm{w}\in\mathbb{R}^p,\ \mathrm{z}\in\mathbb{N}^q} & f(\mathrm{w},\mathrm{z}) = \mathrm{c}^t\mathrm{w} + \mathrm{d}^t\mathrm{z} & \longleftarrow \text{linear} \\ \text{s.t.} & A\mathrm{w} + B\mathrm{z} \leq \mathrm{b} & \longleftarrow \text{linear} \\ & \mathrm{w} \geq 0, \end{cases}$$

for some given $\mathrm{c} \in \mathbb{R}^p, \mathrm{d} \in \mathbb{R}^m, A \in \mathbb{R}^{m\times p}, B \in \mathbb{R}^{m\times q}$ and $\mathrm{d} \in \mathbb{R}^q$.

- A mixed binary linear program is a MILP with $\mathrm{z} \in \{0,1\}^q$ binary.
- When its domain is not empty and bounded, a MILP admit a unique global minimum.

Mixed integer quadratic program (MIQP)

- quadratic cost
- linear constraints
- integer and continuous variables

Definition (mixed integer quadratic program – MIQP)

$$\begin{cases} \min\limits_{\mathrm{x}=(\mathrm{w},\mathrm{z})\in\mathbb{R}^p\times\mathbb{N}^q} & f(\mathrm{x}) = \frac{1}{2}\mathrm{x}^t Q\mathrm{x} + \mathrm{c}^t\mathrm{x} & \longleftarrow \text{quadratic} \\ \text{s.t.} & A\mathrm{x} \leq \mathrm{b} & \longleftarrow \text{linear} \\ & \mathrm{x} \geq 0, & \end{cases}$$

for some given symmetric matrix $Q \in \mathbb{R}^{(p+q)\times(p+q)}$

Mixed integer quadratically constrained quadratic program (MIQCP).

- quadratic cost, quadratic constraints, integer and continuous variables

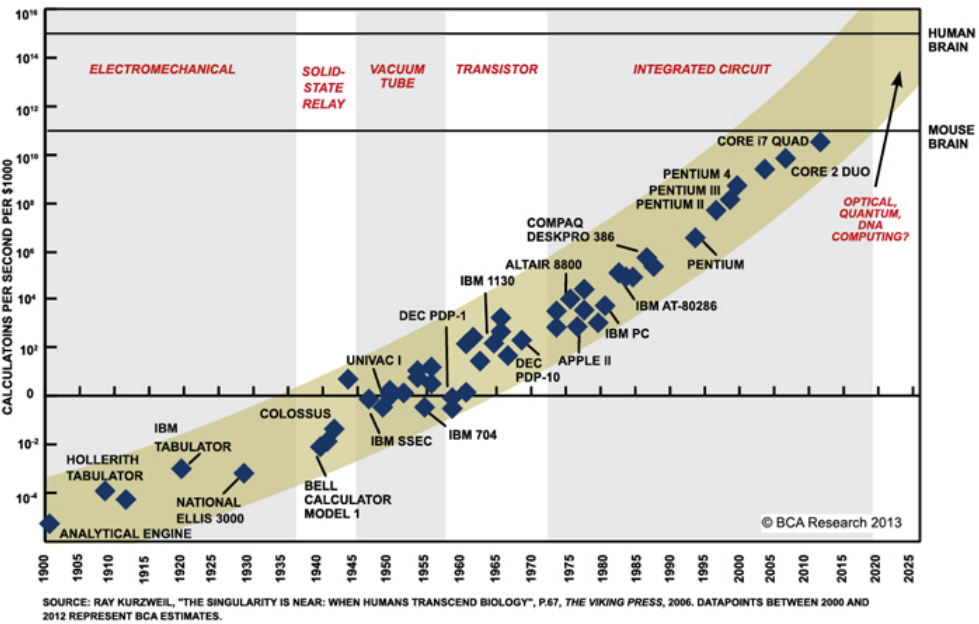
Problems hierarchy

$$\text{MILP } (= \text{MIQP with } Q = 0) \quad \subset \quad \text{MIQP} \quad \subset \quad \text{MIQCP}$$

Progresses in MILP

in 1989

MILP is a powerful modeling tool, "They are, however, theoretically complicated and computationally cumbersome" [?]



from 1996 to 2016 [? ?]

	improvement factor
machine	$\times 2^{10} = 1000 - 1600$
solver	$\times 1000 - 3600$
formulation	???
global	$\times 1 - 5\ 10^6$

a year to solve 10 – 20 years ago $\longrightarrow$ now 30 seconds

"mixed integer linear techniques are nowadays mature, that is fast, robust, and are able to solve problems with up to millions of variables" [?]

Mixed integer software (available with matlab)

Software package	Matlab function
Open source	
GLPK	glpk for mixed integer linear programming
COIN OR	(not matlab yet)
Commercial	
Matlab	intlinprog for mixed integer linear programming
CPLEX	cplexmilp for mixed integer linear programming
	cplexmiqp for mixed integer quadratic programming
	cplexmiqcp for mixed integer quadratically constrained pg
GUROBI	gurobi for MILP, MIQP and MIQCQP
Mosek	mosekopt for MILP, MIQP and MIQCQP
SCIP	opti_scip for MILP, MIQP and MIQCQP

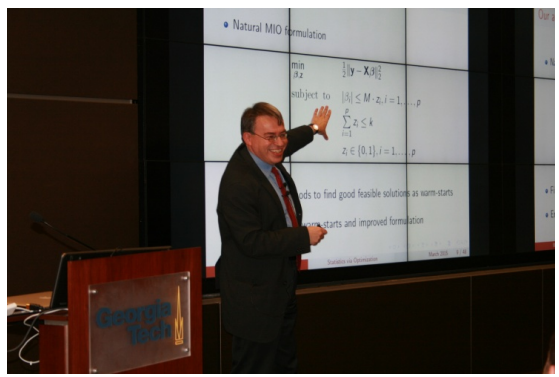
Mixed Integer Linear Programming Benchmark (MIPLIB2010)

recommend CPLEX, GUROBI and Mosek (NOT intlinprog)

http://plato.asu.edu/ftp/milpc.html

So far

- The MIP hierarchy: MILP (linear) $\subset$ MIQP (quadratic) $\subset$ MIQCQP
- use standard MIP software
- the joy of having an exact solution
- formulate ML problems as a MILP or MIQP (if possible)



Bi robust regression

Variable selection AND outlier detection

$$\begin{cases} \min\limits_{\mathrm{w}\in\mathbb{R}^p} & \frac{1}{2}\,\|X\mathrm{w}-\mathrm{y}\|^2 \\ \text{s.t.} & \|w\|_0 \leq k_v \end{cases} \quad \text{and} \quad \begin{cases} \min\limits_{\mathrm{w}\in\mathbb{R}^p,\mathrm{o}\in\mathbb{R}^n} & \frac{1}{2}\,\|X\mathrm{w}+\mathrm{o}-\mathrm{y}\|^2 \\ \text{s.t.} & \|\mathrm{o}\|_0 \leq k_o \end{cases}$$

LS regression with variable selection AND outlier detection

Given k_v the number of variable required and k_o the number of outliers

$$\begin{cases} \min\limits_{w\in\mathbb{R}^p,\mathrm{o}\in\mathbb{R}^n} & \|Xw-y-\mathrm{o}\|^2 \\ \text{s.t.} & \|w\|_0 \leq k_v \\ & \|\mathrm{o}\|_0 \leq k_o. \end{cases} \tag{3}$$

LS with fixed cardinality as a MIQP: the big M constraint

Assuming we know an upper bound M for w

$$\|\mathrm{w}\|_0 \leq k \quad \Leftrightarrow \quad \begin{cases} z_j \in \{0,1\}, \quad j = 1:p \\ \sum_{i=1}^{p} z_j \leq k \\ |w_j| \leq z_j M \end{cases}$$

For useless variables:
$z_j = 0 \quad \Rightarrow \quad w_j = 0$

LS with fixed cardinality as a MIQP [?]

$$\begin{cases} \min\limits_{w \in \mathbb{R}^p, z \in \{0,1\}^p} & \frac{1}{2}\|X\mathrm{w} - \mathrm{y}\|_2^2 \\ \text{s.t.} & \sum_{j=1}^{p} z_j \leq k \\ \text{and} & |w_j| \leq z_j M \qquad j = 1, p \end{cases}$$

Variable selection AND outlier detection as a MILP

$$q \in \{1,2\} \qquad \begin{cases} \min\limits_{w\in\mathbb{R}^p, o\in\mathbb{R}^n} & \|Xw - y - o\|_q^q \\ \text{s.t.} & \|w\|_0 \leq k_v \\ & \|o\|_0 \leq k_o. \end{cases}$$

$q = 1$

$$\begin{cases} \min\limits_{w\in\mathbb{R}^p, o, \varepsilon^+, \varepsilon^- \in\mathbb{R}^n, z\in\{0,1\}^p, t\in\{0,1\}^n} & \sum_{i=1}^{n} \varepsilon_i^+ + \varepsilon_i^- \\ \text{s.t.} & \varepsilon_i^+ - \varepsilon_i^- = x_i^t w + o_i - y_i \quad i = 1, n \\ & \sum_{j=1}^{p} z_j \leq k_v \\ & |w_j| \leq z_j M_v \quad j = 1, p \\ & \sum_{i=1}^{n} (1 - t_i) \leq k_o \\ & |o_i| \leq t_i M_o \quad i = 1, n \\ & 0 \leq \varepsilon_i^+,\ 0 \leq \varepsilon_i^- \quad i = 1, n. \end{cases}$$

LSE with fixed cardinality as a MIQP with SOS constraints

Variable selection: $z_j = 0 \Rightarrow w_j = 0$ either $w_j = 0$ or $1 - z_j = 0$

Special ordered set (SOS) of type 1: at most one variable in the set can take a nonzero value,

$$w_j = 0 \text{ or } 1 - z_j = 0 \quad \Leftrightarrow \quad (w_j, 1 - z_j) : SOS$$

MIQP using special ordered set (SOS) of type 1

$$\begin{cases} \min\limits_{\mathrm{w} \in \mathbb{R}^p, \mathrm{e} \in \mathbb{R}^n, \mathrm{z} \in \{0,1\}^p} & \sum\limits_{i=1}^{n} \frac{1}{2} \left(X_i^t \mathrm{w} - y_i\right)^2 & \longleftarrow \text{ data loss} \\ \text{s.t.} & \sum\limits_{j=1}^{p} z_j \leq k & \longleftarrow \text{ at most } k \text{ non 0 variables} \\ & (w_j, 1 - z_j) : SOS & j = 1, p \end{cases}$$

Variable selection AND outlier detection as a MIQP

$$q \in \{1,2\} \qquad \begin{cases} \min\limits_{w\in\mathbb{R}^p, o\in\mathbb{R}^n} & \|Xw - y - o\|_q^q \\ \text{s.t.} & \|w\|_0 \leq k_v \\ & \|o\|_0 \leq k_o. \end{cases}$$

$q = 2$

$$\begin{cases} \min\limits_{w\in\mathbb{R}^p, o\in\mathbb{R}^n, z\in\{0,1\}^p, t\in\{0,1\}^n} & (y - Xw - o)^t (y - Xw - o) \\ \text{s.t.} & \sum\limits_{j=1}^{p} z_j = k_v \\ & \sum\limits_{i=1}^{n} t_i \leq k_o \\ & (w_j, 1 - z_j) : SOS \qquad j = 1, p \\ & (o_i, 1 - t_i) : SOS \qquad i = 1, n. \end{cases}$$

Balls and Triks: the convex hull of the feasible set

$$Conv\left(\left\{\mathrm{w}\middle|\ |w_j| \le z_j M \text{ and } \sum_{j=1}^{p} z_j \le k\right\}\right) = \left\{\mathrm{w}\middle|\ ||w||_\infty \le M \text{ and } ||\mathrm{w}||_1 \le kM\right\}$$

MIQP: a more structured representation [?]

$$\begin{cases} \min\limits_{\mathrm{w}\in\mathbb{R}^{p},\mathrm{e}\in\mathbb{R}^{n},\mathrm{z}\in\{0,1\}^{p}} & \sum\limits_{i=1}^{n} \frac{1}{2}\left(X_i^t \mathrm{w} - y_i\right)^2 & \longleftarrow \text{ data loss} \\ \text{s.t.} & \sum\limits_{j=1}^{p} z_j \le k & \longleftarrow \text{ at most } k \text{ non 0 variables} \\ & (w_j, 1-z_j) : SOS & j = 1, p \\ & |w_j| \le M_\infty & j = 1, p \\ & \sum\limits_{j=1}^{p} |w_j| \le M_1 & \end{cases}$$

[?] claim: *Adding these bounds typically leads to improved performance of the MIO, especially in delivering lower bound certificates*

Balls and Triks: the convex hull of the feasible set

$$\mathcal{S} = \left\{ \mathrm{w}, \mathrm{o} \mid \sum_{j=1}^{p} z_j \leq k_v,\ |w_j| \leq z_j M_v,\ \sum_{i=1}^{n} (1 - t_i) \leq k_o,\ |\tau_i| \leq t_i M_o \right\},$$

$$Conv(\mathcal{S}) = \left\{ \mathrm{w}, \mathrm{o} \mid ||\mathrm{w}||_\infty \leq M_v\ ||\mathrm{o}||_\infty \leq M_o\ ||\mathrm{w}||_1 \leq k_v M_v,\ ||\mathrm{o}||_1 \leq k_o M_o \right\}$$

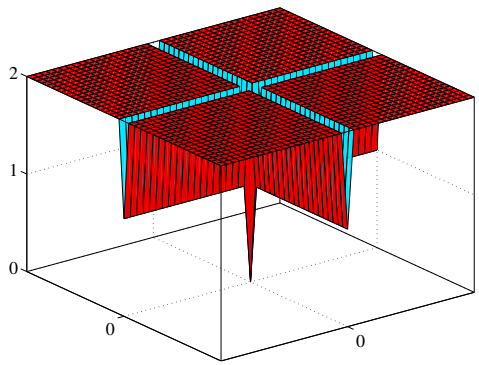
$$\begin{cases} \min\limits_{\mathrm{w},\mathrm{o},\mathrm{z}\in\{0,1\}^p,\mathrm{t}\in\{0,1\}^n} & \frac{1}{q}\|X\mathrm{w} + \mathrm{o} - \mathrm{y}\|_q^q \\ \text{s.t.} & \sum\limits_{j=1}^{p} z_j \leq k_v, \qquad (w_j, 1 - z_j) : SOS \qquad j = 1, p \\ & \sum\limits_{i=1}^{n} (1 - t_i) \leq k_o,\ (\tau_i, 1 - t_i) : SOS \qquad i = 1, n \\ & \|\mathrm{w}\|_1 \leq k_v M_v,\ \ \|\mathrm{w}\|_\infty \leq M_v \\ & \|\mathrm{o}\|_1 \leq k_o M_o,\ \ \|\mathrm{o}\|_\infty \leq M_o, \end{cases} \tag{4}$$

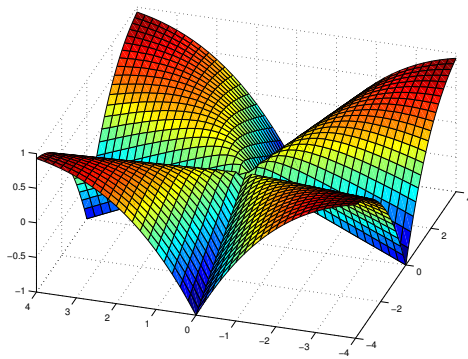
with problem-dependent constants M_v and M_o.

So far...

- birobust regression as a MIP
  - for variable selection AND outlier detection in regression
  - and in quantile regression, SVM, logistic regression
  - reformulation (practical matter)
- efficient software for moderate size problem
- for large size: use first order algorithms

Road map





Variable selection: a specific case with a closed-form solution

Definition (the least square variable selection problem with $X = Id$)

given $k < p$

$$\begin{cases} \min\limits_{\mathrm{u} \in \mathbb{R}^p} & \|\mathrm{u} - \mathrm{w}\|^2 & \longleftarrow \text{fit the data} \\ \text{s.t.} & \|\mathrm{u}\|_0 \leq k & \longleftarrow \text{with k variables} \end{cases}$$

sort $|\mathrm{w}|$: $|w_{(1)}| \geq |w_{(2)}| \geq \ldots |w_{(j)}| \geq \ldots |w_{(p)}|$

Closed-form solution: the hard thresholding operator

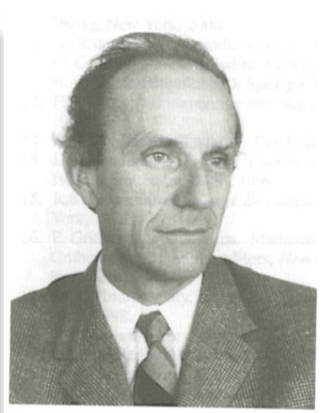
$$u_i^\star = H_k(\mathrm{w}) = \begin{cases} w_j & \text{if } j \in \{(1), \ldots, (k)\} \\ 0 & \text{else} \end{cases}$$

Proximity operator

Definition (Proximity operator [?])

The Proximity operator of a function h is:

$$\begin{array}{rccl} \mathbf{prox}_h : & \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & \mathrm{w} & \longmapsto & \mathbf{prox}_h(\mathrm{w}) = \underset{\mathrm{u}\in\mathbb{R}^p}{\arg\min}\ h(\mathrm{u}) + \frac{1}{2}\|\mathrm{u}-\mathrm{w}\|^2 \end{array}$$



Example

$h(\mathrm{w}) = 0$	$\mathbf{prox}_h(\mathrm{w}) = \mathrm{w}$	
$h(\mathrm{w}) = \rho\mathbf{pen}_\lambda(\mathrm{w})$	$\mathbf{prox}_h(\mathrm{w}) = \mathbf{shr}_{\rho\lambda}(\mathrm{w})$	shrinkage
$h(\mathrm{w}) = \mathbb{I}_C(\mathrm{w})$	$\mathbf{prox}_h(\mathrm{w}) = \underset{\mathrm{u}\in C}{\arg\min}\ \frac{1}{2}\|\mathrm{u}-\mathrm{w}\|^2$	projection

The proximity operator as a projection

$$\mathbf{prox}_{\mathbb{I}_{\{\|\mathrm{w}\|_0\leq k\}}}(\mathrm{w}) = \underset{\|\mathrm{u}\|_0\leq k}{\arg\min}\ \frac{1}{2}\|\mathrm{u}-\mathrm{w}\|^2 = H_k(\mathrm{w}) = \begin{cases} w_i & \text{if } i \in \{(1),\ldots,(k)\} \\ 0 & \text{else} \end{cases}$$

The projected gradient (L_0 projection or proximal)

$$\text{for solving} \quad \begin{cases} \min\limits_{w \in \mathbb{R}^p} & \frac{1}{2}\|X\mathrm{w} - \mathrm{y}\|^2 \\ \text{s.t.} & \|\mathrm{w}\|_0 \leq k_v \end{cases}$$

Algorithm 1 L_0 gradient projection algorithm [?]

Data: $X, \mathrm{y}, \mathrm{w}$ initialization
Result: w
while *not converged* **do**
  $g \leftarrow \nabla g(\mathrm{w}) = X^\top(X\mathrm{w} - y)$, the gradient
  $\rho \leftarrow$choose a stepsize
  $d \leftarrow \mathrm{w} - \rho g$, forward (explicit)
  $\mathrm{w} \leftarrow H_k(d)$, the projection–proximal step, backward (implicit)
end

if $\varepsilon \leq \rho \leq \frac{1}{\|X^\top X\|}$, it converges towards a local minimum [?] since its objective function satisfies the Kurdyka-Lojasiewicz inequality.

Proximal alternating linearized minimization (PALM)

$$\begin{cases} \min\limits_{w\in\mathbb{R}^p, o\in\mathbb{R}^n} & \frac{1}{2}\|Xw + o - y\|^2 \\ \text{s.t.} & \|w\|_0 \le k_v \\ & \|o\|_0 \le k_o \end{cases}$$

given o

$$\begin{cases} \min\limits_{w\in\mathbb{R}^p} & \frac{1}{2}\|Xw + o - y\|^2 \\ \text{s.t.} & \|w\|_0 \le k_v \end{cases}$$

given w

$$\begin{cases} \min\limits_{o\in\mathbb{R}^n} & \frac{1}{2}\|o - (y - Xw)\|^2 \\ \text{s.t.} & \|o\|_0 \le k_o \end{cases}$$

Algorithm 2 Proximal alternating linearized minimization (PALM) [?]

Data: X, y initialization $w, o = 0$

Result: w, o

while *not converged* **do**

$d \leftarrow w - \rho_v X^\top (Xw + o - y)$, variable selection

$w \leftarrow H_{k_v}(d)$,

$\delta \leftarrow o - \rho_o (Xw + o - y)$, eliminating outliers

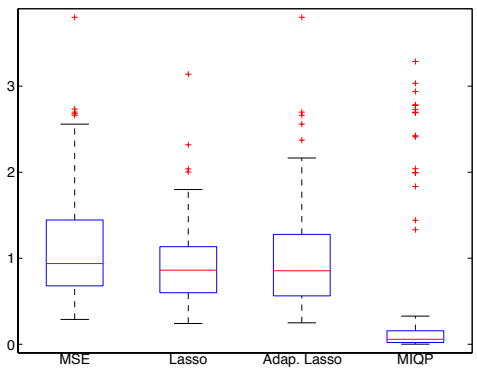
$o \leftarrow H_{k_o}(\delta)$,

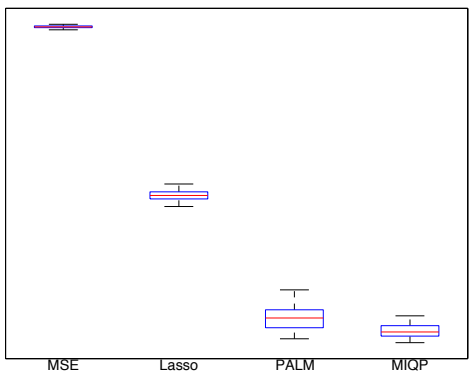
end

Prox summary

- PALM is fast and scalable
- convergence profs towards a local minimum
- improvement:
  - accelerations: FISTA and others
  - Newton proximal
  - more improvement with randomization

Road map





Combine the best of the two worlds

Combine the best of the two worlds [?]

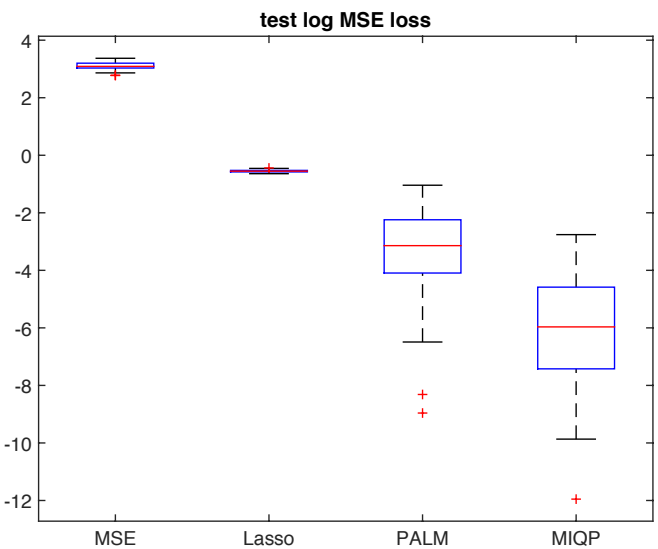
1. $(\mathrm{w}, \mathrm{o}) \leftarrow$ PALM alternating proximal gradient method
2. use w and o as a warm start for MIP (with Cplex)
   - $(\mathrm{w}, \mathrm{o}) \leftarrow$ Polish coefficients on the active set
   - initialize the constants M_v, M_o

Experimental setup

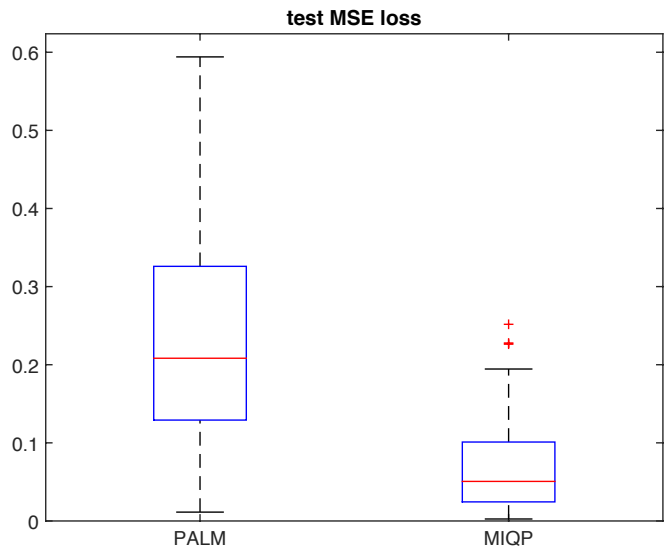
- My mac
- Matlab
- Cplex 12.6.1 (`cplexmiqp`)
- time out = 5 min

Variable selection AND outlier detection on a toy dataset

- $\mathrm{y} = X\mathrm{w} + \mathrm{o} + \varepsilon$
- $n = 300$ observations with $p = 25$ variables
- linear model with ε a centered gaussian noise with SNR ≈ 1
- $k_o = 50$ outliers and $k_v = 5$ non zeros variables
- 100 repetions



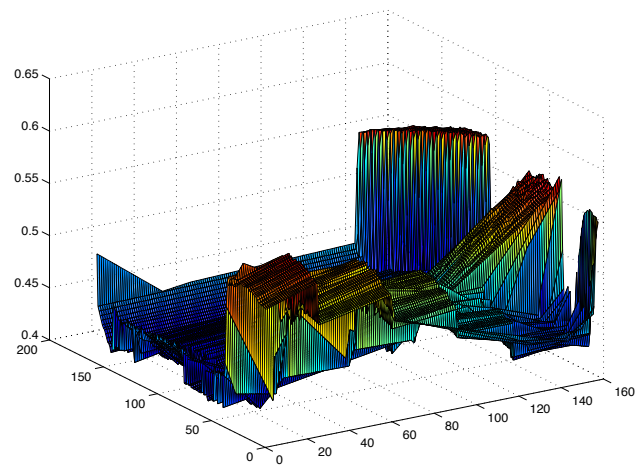
(log) performances



zoom

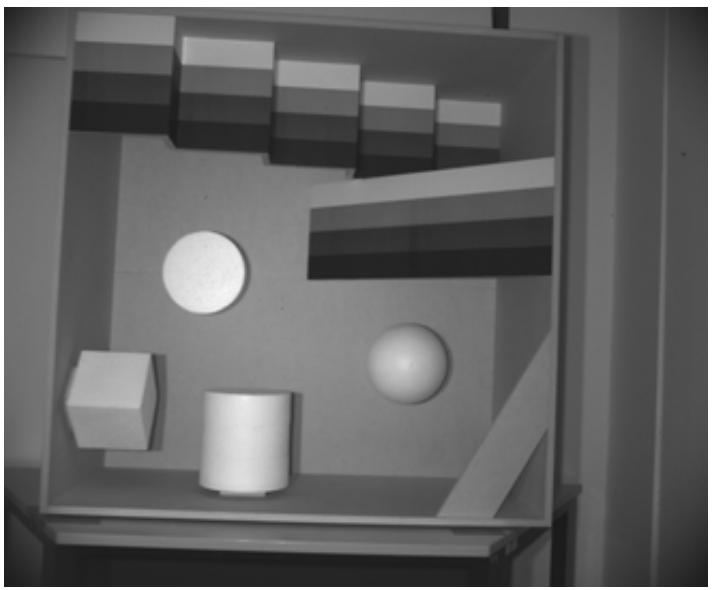
Road map

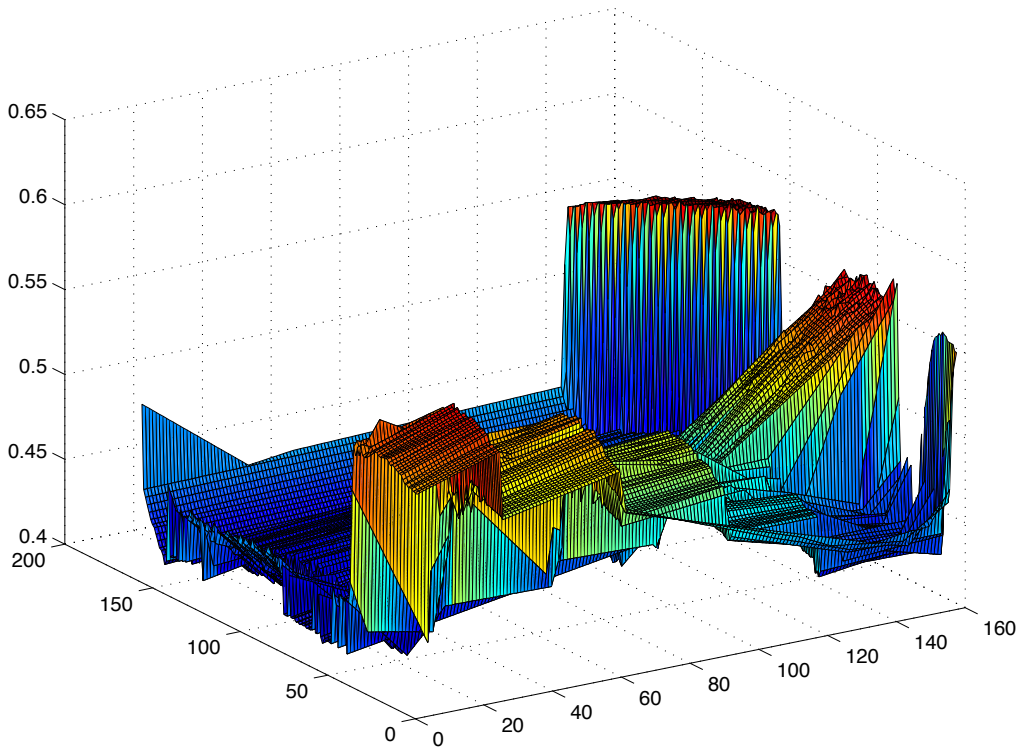
1. Examples of combinatorial problems in machine learning
2. Mixed integer (binary) programming (MIP)
3. L_0 proximal algorithm
4. Implementation
5. MIP for image processing



A Second Order Total Variation MIP Model for Image Segmentation Using the L_0 Norm
(*on going work*)

The problem: deepth estimation





Fundamental hypothesis

piecewise linear model

The second order TV Potts model in the 1d case

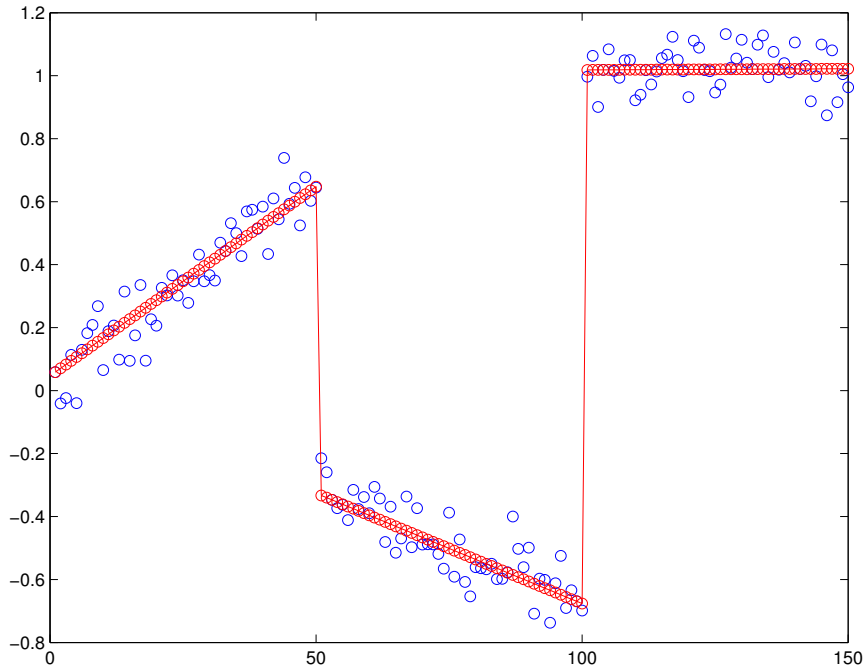
The objective: given n observations $(z_1, \ldots, z_n)$

Retrieve the best linear fit with k discontinuities

$$\begin{array}{ll} \min\limits_{\mathbf{w}\in\mathbb{R}^n} & \|\mathbf{w}-\mathbf{z}\|_1 \\ \text{s.t.} & \|\nabla_x^2 \mathbf{w}\|_0 \leq k \end{array}$$

piece wise linear model

$$\begin{array}{c} w_i = a_\ell i + b_\ell \\ \ell = 1, \ldots, k^* + 1; \end{array}$$

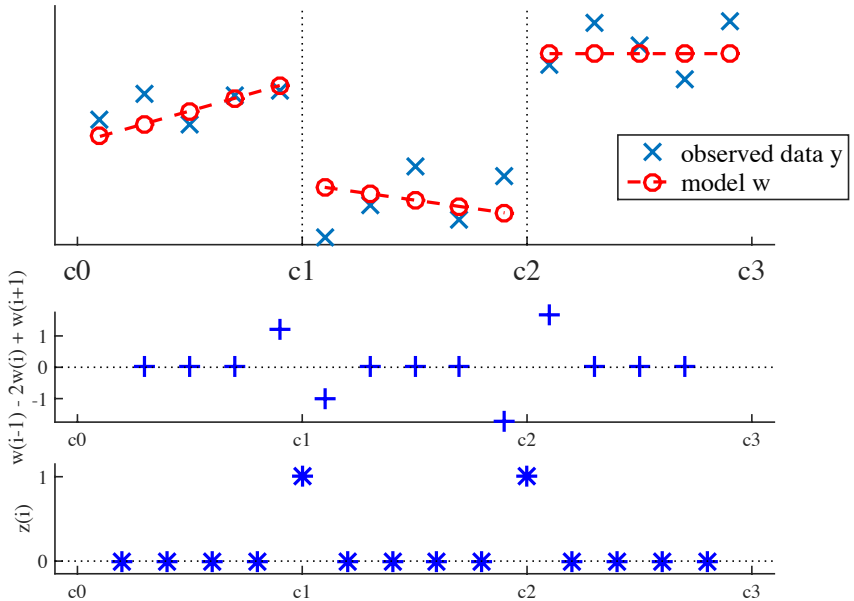


$\|\nabla_x^2 \mathbf{w}\|_0 \leq k$ with k small

- $\|\nabla_x^2 \mathbf{w}\|_0 = 0$ impose linear model
- $\|\nabla_x^2 \mathbf{w}\|_0 \neq 0$ allow discontinuity

The second order TV Potts model as a MIP

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^n} \quad & \|\mathbf{w}-\mathbf{z}\|_1 \\ \text{s.t.} \quad & \|\nabla_x^2 \mathbf{w}\|_0 \le k \end{aligned}$$



$$\nabla_x^2 \mathbf{w} \propto w_{i-1} - 2w_i + w_{i+1}$$

$$\begin{cases} w_{i-1} - 2w_i + w_{i+1} = 0 \quad \Rightarrow \quad x_i = 1, \quad i = 2,\ldots,n-1, \\ w_{i-1} - 2w_i + w_{i+1} \neq 0 \quad \Rightarrow \quad x_i = 0, \quad i = 2,\ldots,n-1 \end{cases}$$

For a given M large enough (the big M trick)

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^n, \mathbf{x}\in\{0,1\}^{n-1}} \quad & \sum_{i=1}^{n} |w_i - z_i| \\ \text{s.t.} \quad & |w_{i-1} - 2w_i + w_{i+1}| \le M(x_{i-1} + x_i), \quad i = 2,\ldots,n-1, \\ & \sum_{i=1}^{n-1} x_i \le k. \end{aligned}$$

Eliminating the absolute values: positive and negative parts

$$\min_{\mathbf{w}\in\mathbb{R}^{n},\mathbf{x}\in\{0,1\}^{n-1}} \quad \sum_{i=1}^{n} |w_i - z_i|$$

$$\text{s.t.} \quad |w_{i-1} - 2w_i + w_{i+1}| \le M(x_{i-1} + x_i), \quad i = 2, \ldots, n-1,$$

$$\sum_{i=1}^{n-1} x_i \le k.$$

Positive and negative parts: $z_+, z_- \ge 0$

$$z = z_+ - z_- \qquad \text{and} \qquad |z| = z_+ + z_-$$

$$\underset{\mathbf{w},\varepsilon^+,\varepsilon^-\in\mathbb{R}^{n},\mathbf{x}\in\{0,1\}^{n-1}}{\text{minimize}} \quad \sum_{i=1}^{n}(\varepsilon_i^+ + \varepsilon_i^-)$$

$$\text{s.t.} \quad \mathbf{w} - \mathbf{z} = \varepsilon^+ - \varepsilon^-$$

$$w_{i-1} - 2w_i + w_{i+1} \le M(x_{i-1} + x_i), \quad i = 2, \ldots, n-1$$

$$-w_{i-1} + 2w_i - w_{i+1} \le M(x_{i-1} + x_i), \quad i = 2, \ldots, n-1$$

$$\sum_{i=1}^{n-1} x_i \le k \qquad\qquad 0 \le \varepsilon^+,\ 0 \le \varepsilon^-$$

How to accelerate the solver?

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^{n},\mathbf{x}\in\{0,1\}^{n-1}} \quad & \sum_{i=1}^{n} |w_i - z_i| \\ \text{s.t.} \quad & |w_{i-1} - 2w_i + w_{i+1}| \le M(x_{i-1} + x_i), \quad i = 2, \ldots, n-1, \\ & \sum_{i=1}^{n-1} x_i \le k. \end{aligned}$$

Initialization: use first order algorithm (ADMM, proximal, greedy...)

- initialize $\mathbf{w}$ and $\mathbf{x}$
- initialize parameter M

Stronger constraints

- based on the convex hull of the actual constraints
- typically: $\|D_n w\|_1 \le k$
- local implied bound cuts to treat the big-M [?]

[?] claim: *Adding these bounds typically leads to improved performance of the MIO, especially in delivering lower bound certificates*

First order algorithm :toward a local minimum

$$D_n = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ & & & \ddots & & \\ 0 & \dots & 0 & 1 & -2 & 1 \end{pmatrix} \in \mathbb{R}^{n-2 \times n}$$

$$\begin{cases} \min\limits_{\mathbf{w} \in \mathbb{R}^n} & \|\mathbf{w} - \mathbf{z}\|_1 \\ \text{s.t.} & \|D_n \mathbf{w}\|_0 \leq k \end{cases} \Rightarrow \begin{cases} \min\limits_{\mathbf{w} \in \mathbb{R}^n, \mathbf{r} \in \mathbb{R}^{n-2}} & \|\mathbf{w} - \mathbf{z}\|_1 \\ \text{s.t.} & \|\mathbf{r}\|_0 \leq k \end{cases} \quad \text{and} \quad \mathbf{r} = D_n \mathbf{w}$$

The augmented Lagrangian with $\lambda > 0$

$$L(\mathbf{w}, \mathbf{r}, \Lambda) = \|\mathbf{w} - \mathbf{z}\|_1 + \Lambda^t(\mathbf{r} - D_n \mathbf{w}) + \frac{\lambda}{2}\|\mathbf{r} - D_n \mathbf{w}\|^2 + I_{\{\|\mathbf{r}\|_0 \leq k\}}$$

The ADMM algorithm:

$$\begin{aligned} \mathbf{w}^{k+1} &= \arg\min_{\mathbf{w}} L(\mathbf{w}^k, \mathbf{r}^k, \Lambda^k) \\ \mathbf{r}^{k+1} &= \arg\min_{\mathbf{r}} L(\mathbf{w}^{k+1}, \mathbf{r}^k, \Lambda^k) \\ \Lambda^{k+1} &= \Lambda^k + \rho(D_n \mathbf{w}^{k+1} - \mathbf{r}^{k+1}) \end{aligned}$$

Combine the best of the two worlds

The problem:

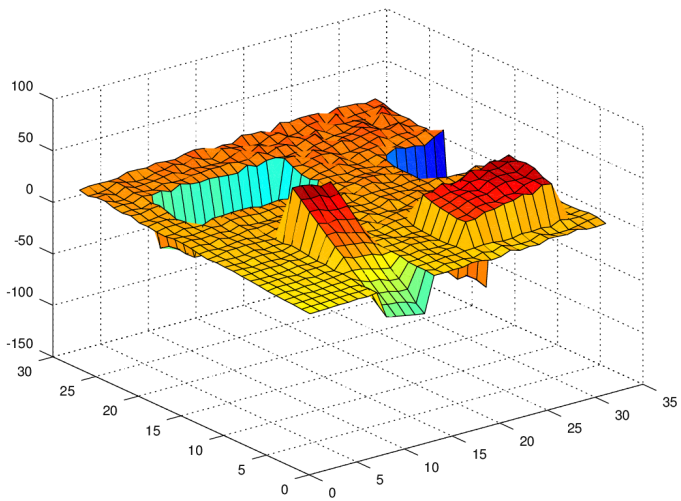
$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^{n},\mathbf{x}\in\{0,1\}^{n-1}} \quad & \sum_{i=1}^{n} |w_i - y_i| \\ \text{s.t.} \quad & |w_{i-1} - 2w_i + w_{i+1}| \leq M(x_{i-1} + x_i), \quad i = 2, \ldots, n-1, \\ & \sum_{i=1}^{n-1} x_i \leq k. \end{aligned}$$

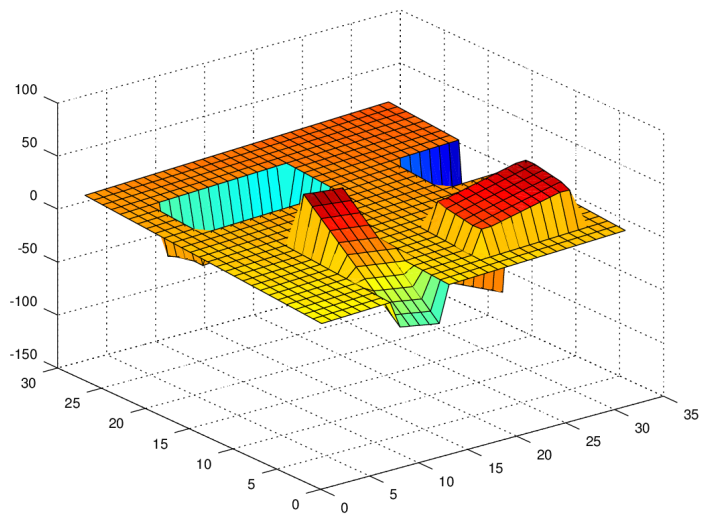
The proposed solution [?]

1. $(\mathbf{w}, \mathbf{x}) \leftarrow$ ADMM alternating proximal gradient method
2. use $\mathbf{w}$ and $\mathbf{x}$ as a warm start for MILP (with Cplex)
   - $(\mathbf{w}, \mathbf{x}) \leftarrow$ Polish coefficients on the active set
   - initialize the constant M

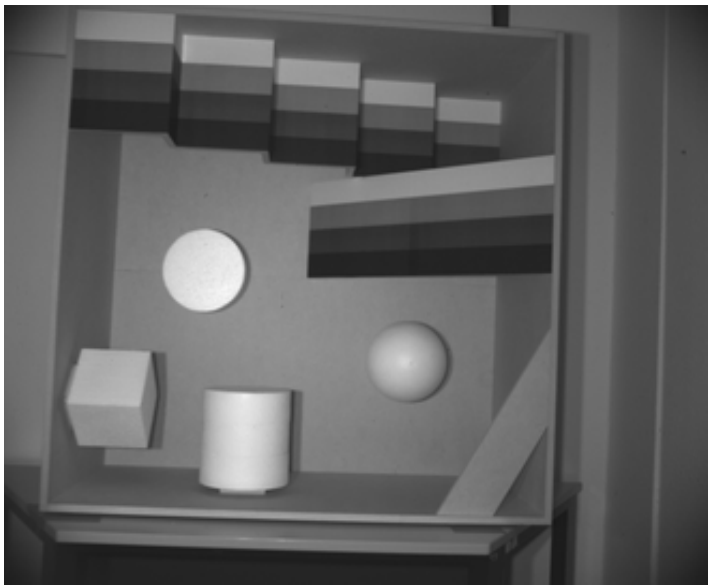
Road map

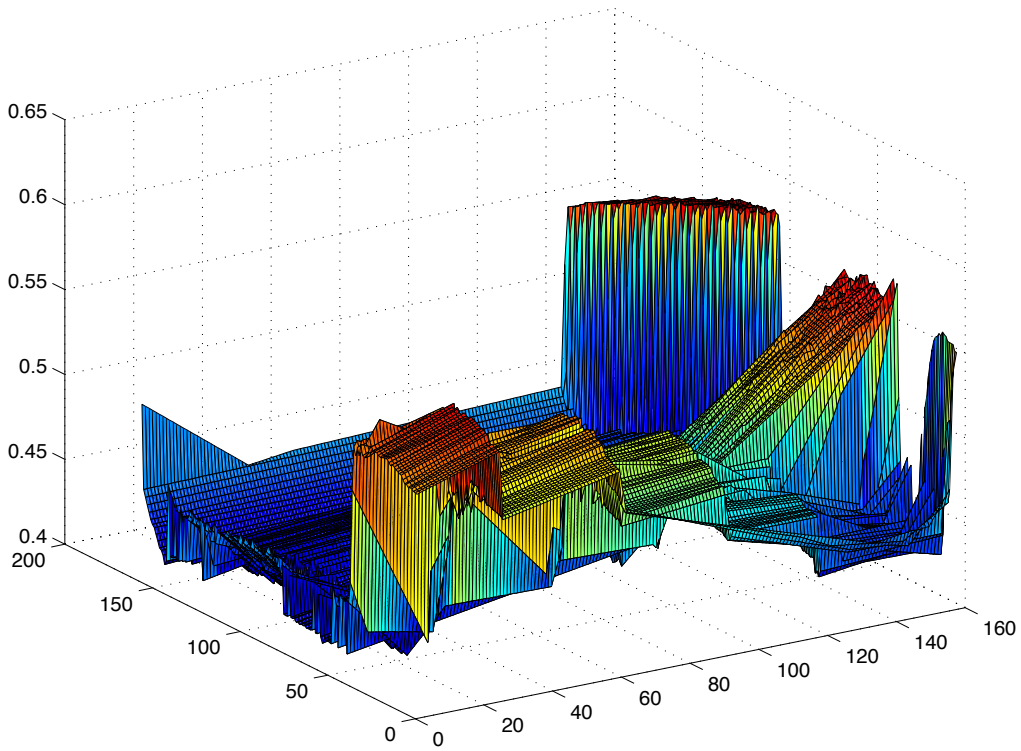
1. Examples of combinatorial problems in machine learning
2. Mixed integer (binary) programming (MIP)
3. L_0 proximal algorithm
4. Implementation
5. MIP for image processing





The problem: deepth estimation based on image Z



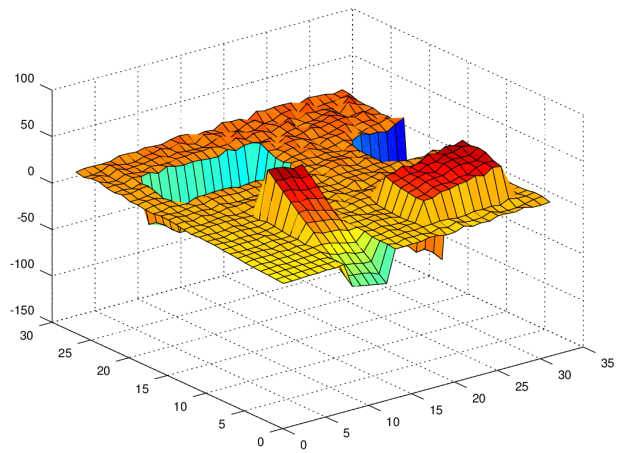


Fundamental hypothesis

piecewise linear model

ℓ_0 constraints on row and columns

$$\begin{aligned} \min_{W \in R^{m \times n}} \quad & \|W - Z\|_1 \\ \text{s.t.} \quad & \|\nabla_x^2 W\|_0 \leq k_{r_i}, \quad \forall i = 1, \ldots, n, \\ & \|\nabla_y^2 W\|_0 \leq k_{c_j}, \quad \forall j = 1, \ldots, m. \end{aligned}$$



Fundamental hypothesis

piecewise 2d linear model

the problem as a MILP

$$\min_{W \in R^{m \times n}} \quad \|W - Z\|_1$$

$$\text{s.t.} \quad \|\nabla_x^2 W\|_0 \leq k_{r_i}, \quad \forall i = 1, \ldots, n,$$

$$\|\nabla_y^2 W\|_0 \leq k_{c_j}, \quad \forall j = 1, \ldots, m.$$

$$\underset{W \in \mathbb{R}^{n \times m}}{\text{minimize}} \quad \sum_{i=1}^{n} \sum_{j=1}^{m} |W_{ij} - Z_{ij}|$$

$x \in \{0,1\}^{n \times (m-1)}$ row

$y \in \{0,1\}^{(n-1) \times m}$ column

$$\text{s.t.} \quad |W_{i,j+1} + W_{i,j-1} - 2W_{ij}| \leq M_{r_i}(x_{i,j-1} + x_{ij}), \quad i = 1, n \quad j = 2, m-1$$

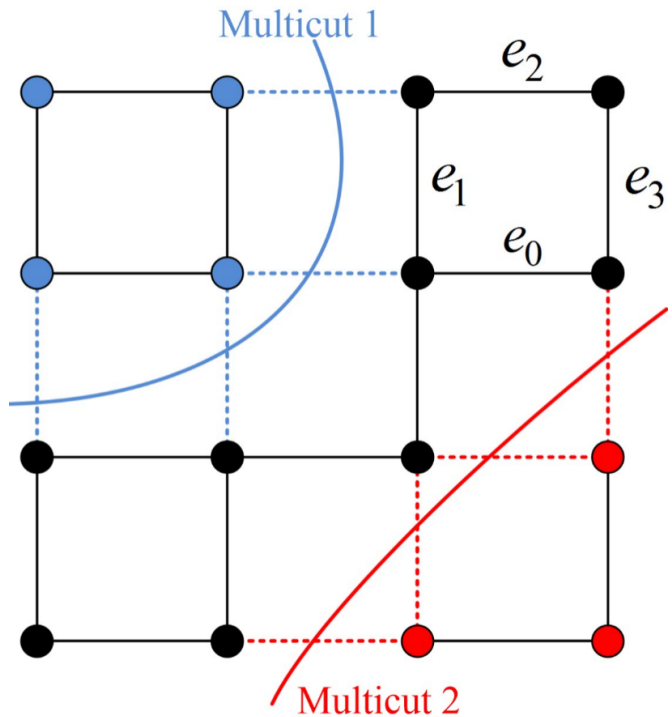
$$|W_{i+1,j} + W_{i-1,j} - 2W_{ij}| \leq M_{c_j}(y_{i-1,j} + y_{ij}), \quad j = 1, m \quad i = 2, n-1$$

$$\sum_{j=2}^{m-1} x_{ij} \leq k_{r_i}, \quad i = 1, , n$$

$$\sum_{i=2}^{n-1} y_{ij} \leq k_{c_j}, \quad j = 1, , m,$$

A strongest formulation with multi-cut constraints

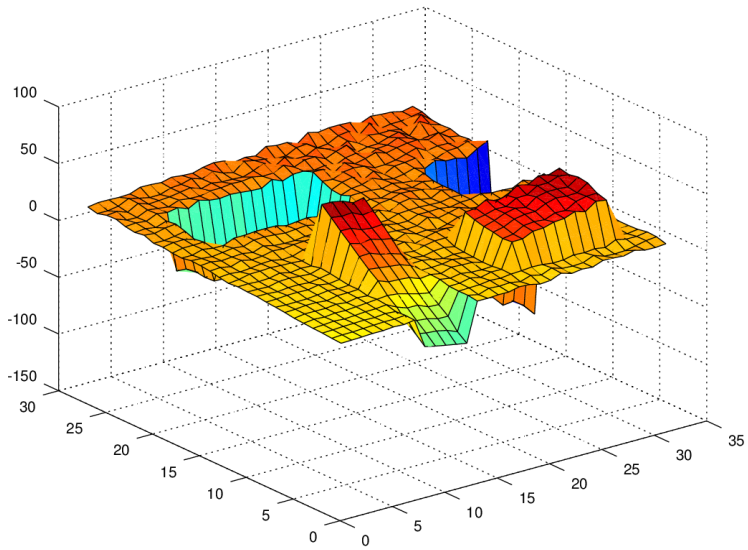
row cut: $x \in \{0,1\}^{n\times(m-1)}$ column cut: $y \in \{0,1\}^{(n-1)\times m}$

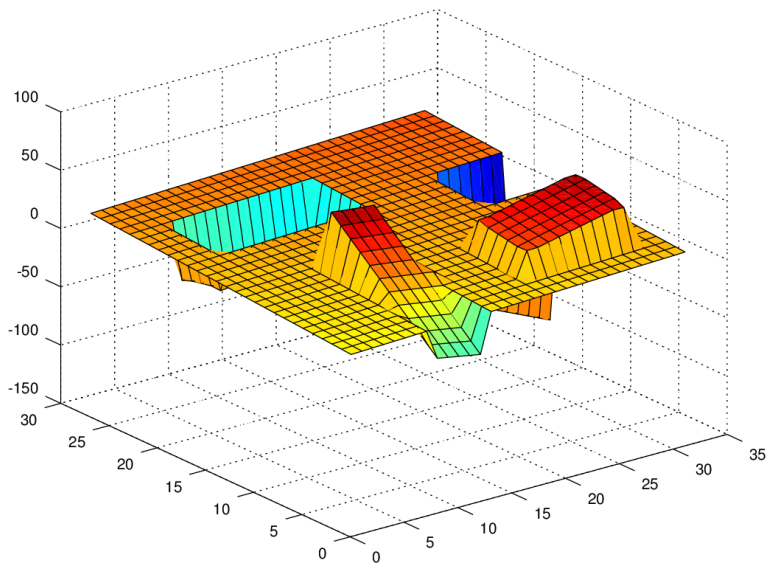


$4 \times (n-2)^2$ additional constraints

$$y_1 + x_2 + y_3 \geq x_0, \qquad x_0 + x_2 + y_3 \geq y_1,$$
$$x_0 + y_1 + y_3 \geq x_2, \qquad x_0 + y_1 + x_2 \geq y_3.$$

The matlab simulation






```
function [ W,X,Y,fval ] = Snd_Order_TV_MILP( Z,kr,kc,Tmax,disp )

Z           28x28   = 784

Reduced MIP has 5283 rows, 4541 columns, and 22807 nonzeros.
Reduced MIP has 1104 binaries, 0 generals, 0 SOSs, and 0 indicators.
```

10 sec to find the minimum

more than 10 minutes to prove it

can be reduced [?]

Road map (done)

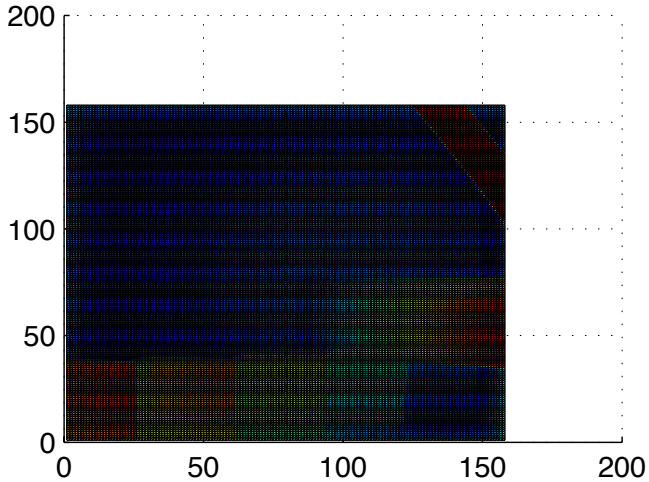
1. Examples of combinatorial problems in machine learning
2. Mixed integer (binary) programming (MIP)
3. L_0 proximal algorithm
4. Implementation
5. MIP for image processing

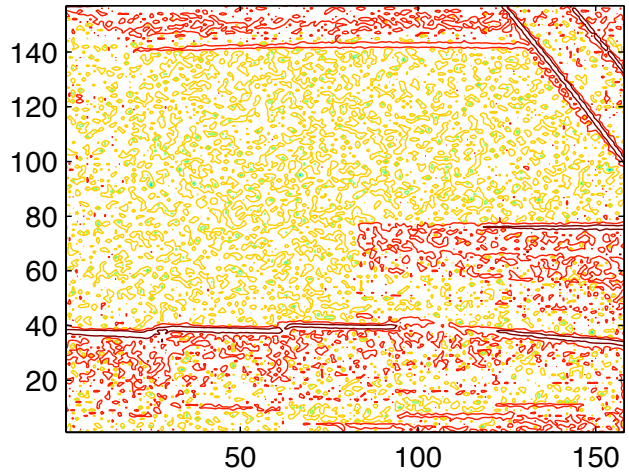
Conclusion

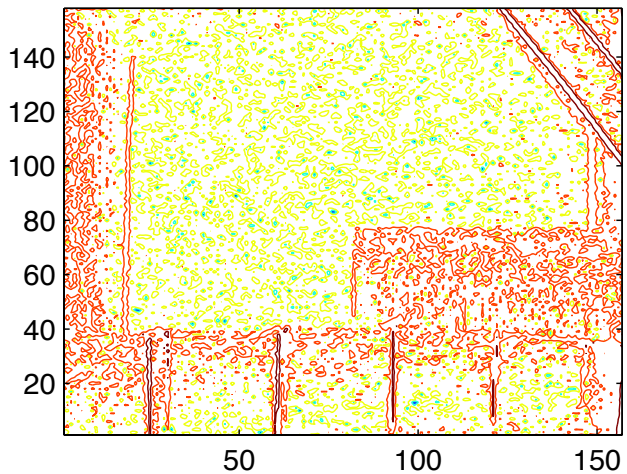
- Machine learning with MIP [?]

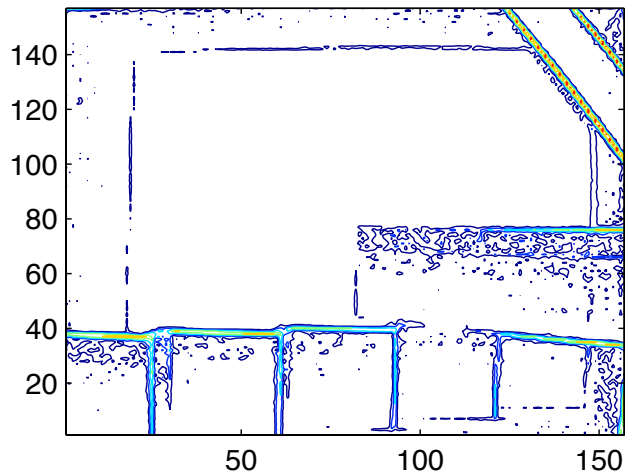
pros	cons
it works	it does not scale
global optimum	only linear or quadratic
flexible	show some instability
that is what we want to do	it's not what we want to do

- Future work
  - efficient generic solver
  - efficient implementation: parallelization, randomisation, GPU
  - efficient hyper parameter calibration


200
150
100
50
0
0
50
100
150
200


140
120
100
80
60
40
20
50
100
150


140
120
100
80
60
40
20
50
100
150


140
120
100
80
60
40
20
50
100
150