Sparse Regularization by Evolving the ℓ_1 Subgradient

Wotao Yin

Department of Mathematics, UCLA

joint with UCLA: Ming Yan and Stanley Osher

PKU: Yuan Yao, Feng Ruan, Jiechao Xiong

Sparse recovery

- **Goal:** recover a **sparse vector** $u \in \mathbb{R}^n$ from **noisy** measurements

$$b = Au + \omega.$$

Sparse recovery

- **Goal:** recover a **sparse vector** $u \in \mathbb{R}^n$ from **noisy** measurements

$$b = Au + \omega.$$

- Given A and b, we have **two tasks:**

1. **variable/predictor selection**: find the support of u
2. **estimation**: predict the values of u

- Largely many applications and several existing approaches

ℓ_1 subgradient

- **Proposed method:** variable selection based on ℓ_1-subgradient p
- Subdifferential of convex function f

$$\partial f(x) = \{p : f(y) \geq f(x) + \langle p, y - x \rangle, \ \forall y \in \mathrm{dom} f\}.$$

$p \in \partial f(x)$ is a subgradient of f at x.

ℓ_1 subgradient

- **Proposed method:** variable selection based on ℓ_1-subgradient p
- Subdifferential of convex function f

$$\partial f(x) = \{p : f(y) \geq f(x) + \langle p, y - x\rangle,\ \forall y \in \mathrm{dom} f\}.$$

$p \in \partial f(x)$ is a subgradient of f at x.

- Subdifferential of $|\cdot|$:

$$\partial|x| = \begin{cases} \{1\}, & x > 0; \\ [-1, 1], & x = 0; \\ \{-1\}, & x < 0. \end{cases}$$

$\implies$ given that $p \in \partial|x|$, then

$$x \begin{cases} \geq 0, & \text{if } p = 1; \\ = 0, & \text{if } p \in (-1, 1); \\ \leq 0, & \text{if } p = -1. \end{cases}$$

- ℓ_1 subdifferential:

$$\partial\|u\|_1 = \partial|u_1| \times \cdots \partial|u_n|.$$

$\implies$ given that $p \in \partial\|u\|_1$, then

$$u_i \begin{cases} \geq 0, & \text{if } p_i = 1; \\ = 0, & \text{if } p_i \in (-1, 1); \\ \leq 0, & \text{if } p_i = -1. \end{cases}$$

- $u_i = \pm 1 \implies u_i$ can be nonzero.
- we select predictors by computing p.

Sparse variable selection

- Suppose $p \in \partial\|u\|_1$

$$u \in \mathbb{R}^n \text{ is sparse} \iff \text{few } p_i = \pm 1$$

- Assume A is short and wide (few rows and more columns)
- $p \in \partial\|u\|_1 \cap \mathcal{R}(A^T) \implies u$ tends to sparse
- Subgaussian random A of appropriate size $\implies$ sparse u w.h.p.

Data fitting

We shall compute p such that

- sparsity: $p \in \partial\|u\|_1 \cap \mathcal{R}(A^T)$
- fitting: $\|Au - b\|_2$ is small

Data fitting

We shall compute p such that

- sparsity: $p \in \partial\|u\|_1 \cap \mathcal{R}(A^T)$
- fitting: $\|Au - b\|_2$ is small

Proposed system:

$$\dot{p}_+(t) = A^*(b - Au(t)), \tag{1a}$$

$$p(t) \in \partial\|u(t)\|_1. \tag{1b}$$

Initial solution: $p(0) = 0$, $u(0) = 0$.

Notation:

- $\dot{p}_+(t)$: right derivative of $p(t)$
- $A^* = \frac{1}{m}A^T$: normalized adjoint
- $\partial\|\cdot\|_1$: ℓ_1 subdifferential

Known as inverse-scale space (ISS) with total variation

Toy example

- Single real measurement

$$b = \mathbf{a}^T u + \epsilon \in \mathbb{R}$$

Suppose $|a_1| > |a_2|, \ldots, |a_n|$ w.o.l.g.

Toy example

- Single real measurement

$$b = \mathbf{a}^T u + \epsilon \in \mathbb{R}$$

Suppose $|a_1| > |a_2|, \ldots, |a_n|$ w.o.l.g.

- Zero initial condition $\implies$

$$\dot{p}_+(t) = \mathbf{a}(b - \mathbf{a}^T 0) = b\mathbf{a}$$

$\implies$

$$p(t) = t \cdot (b\mathbf{a}).$$

Toy example

- Single real measurement

$$b = \mathbf{a}^T u + \epsilon \in \mathbb{R}$$

Suppose $|a_1| > |a_2|, \ldots, |a_n|$ w.o.l.g.

- Zero initial condition $\implies$

$$\dot{p}_+(t) = \mathbf{a}(b - \mathbf{a}^T 0) = b\mathbf{a}$$

$\implies$

$$p(t) = t \cdot (b\mathbf{a}).$$

- At time $t_1 = |ba_1|^{-1}$,

$$p_1(t_1) = \mathrm{sign}(ba_1), \quad p_2(t_1), \ldots, p_n(t_1) \in (-1, 1).$$

Hence, $u_1(t_1)$ can be nonzero.

Under technical assumptions:

- p is right continuously differentiable, and
- u is right continuous,

$u(t_1)$ must be the solution to

$$\underset{u}{\text{minimize}}\, \|\mathbf{a}^T u - b\|_2^2 \quad \text{s.t. } p_1(t_1) \cdot u_1 \geq 0,\ u_2 = \cdots = u_n = 0.$$

Under technical assumptions:

- p is right continuously differentiable, and
- u is right continuous,

$u(t_1)$ must be the solution to

$$\underset{u}{\text{minimize}}\, \|\mathbf{a}^T u - b\|_2^2 \quad \text{s.t. } p_1(t_1) \cdot u_1 \geq 0,\ u_2 = \cdots = u_n = 0.$$

$\Longrightarrow$

$$u_1(t_1) = \frac{b}{a_1}, \quad u_2(t_1) = \cdots = u_n(t_1) = 0.$$

Easy to verify

$$p(t_1) \in \partial \|u(t_1)\|_1.$$

Under technical assumptions:

- p is right continuously differentiable, and
- u is right continuous,

$u(t_1)$ must be the solution to

$$\underset{u}{\text{minimize}}\,\|\mathbf{a}^T u - b\|_2^2 \quad \text{s.t. } p_1(t_1)\cdot u_1 \geq 0,\ u_2 = \cdots = u_n = 0.$$

$\Longrightarrow$

$$u_1(t_1) = \frac{b}{a_1}, \quad u_2(t_1) = \cdots = u_n(t_1) = 0.$$

Easy to verify

$$p(t_1) \in \partial\|u(t_1)\|_1.$$

For $t > t_1$, $p(t) = p(t_1)$ and $u(t) = u(t_1)$ stay constant

General case

Theorem

The solution path to

$$\dot{p}_+(t) = A^*(b - Au(t)), \quad p(t) \in \partial\|u(t)\|_1$$

with initial conditions $t_0 = 0$, $p(0) = 0$, $u(0) = 0$, *is uniquely given by:*

- *for* $k = 1, 2, \ldots, K$

General case

Theorem

The solution path to

$$\dot{p}_+(t) = A^*(b - Au(t)), \quad p(t) \in \partial\|u(t)\|_1$$

with initial conditions $t_0 = 0$, $p(0) = 0$, $u(0) = 0$, *is uniquely given by:*

- *for* $k = 1, 2, \ldots, K$
  - $p(t)$ *is piece-wise linear*

$$p(t) = p(t_{k-1}) + (t - t_{k-1})A^*(b - Au(t_{k-1})), \quad t \in [t_{k-1}, t_k],$$

where

$$t_k := \sup\{t > t_{k-1} : p(t) \in \|u(t_{k-1})\|_1\}.$$

General case

Theorem

The solution path to

$$\dot{p}_+(t) = A^*(b - Au(t)), \quad p(t) \in \partial\|u(t)\|_1$$

with initial conditions $t_0 = 0$, $p(0) = 0$, $u(0) = 0$, *is uniquely given by:*

- *for* $k = 1, 2, \ldots, K$
  - $p(t)$ *is piece-wise linear*

    $$p(t) = p(t_{k-1}) + (t - t_{k-1})A^*(b - Au(t_{k-1})), \quad t \in [t_{k-1}, t_k],$$

    where

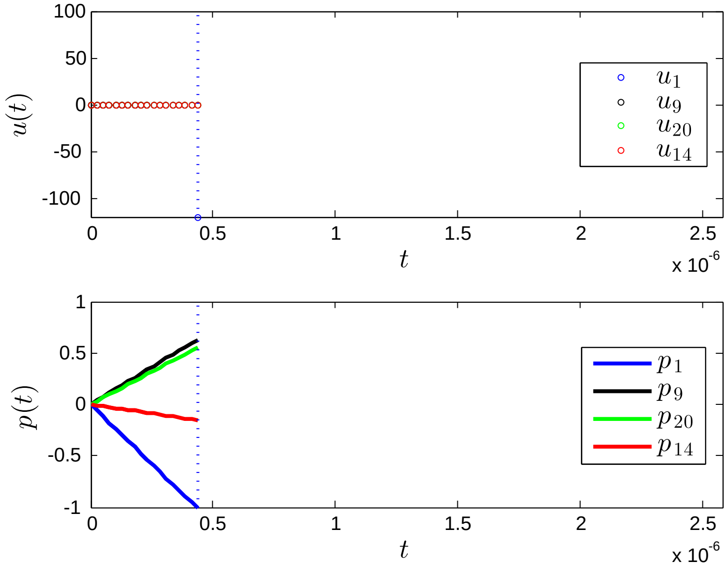
    $$t_k := \sup\{t > t_{k-1} : p(t) \in \|u(t_{k-1})\|_1\}.$$

  - $u(t) = u(t_{k-1})$ *for* $t \in [t_{k-1}, t_k)$;

    if $t_k \neq \infty$, *compute*

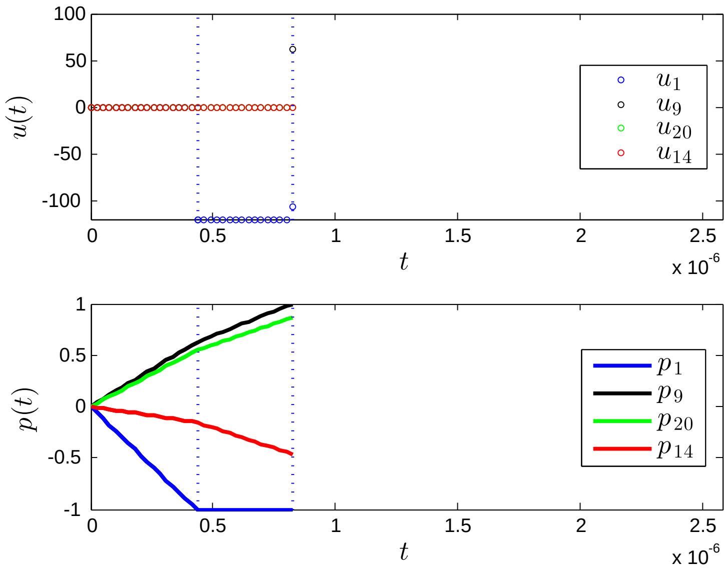
    $$u(t_k) = \arg\min_u \|Au - b\|_2^2 \quad \textit{s.t. } u_i \begin{cases} \geq 0, & p_i(t_k) = 1, \\ = 0, & p_i(t_k) \in (-1, 1), \\ \leq 0, & p_i(t_k) = -1. \end{cases}$$

Example



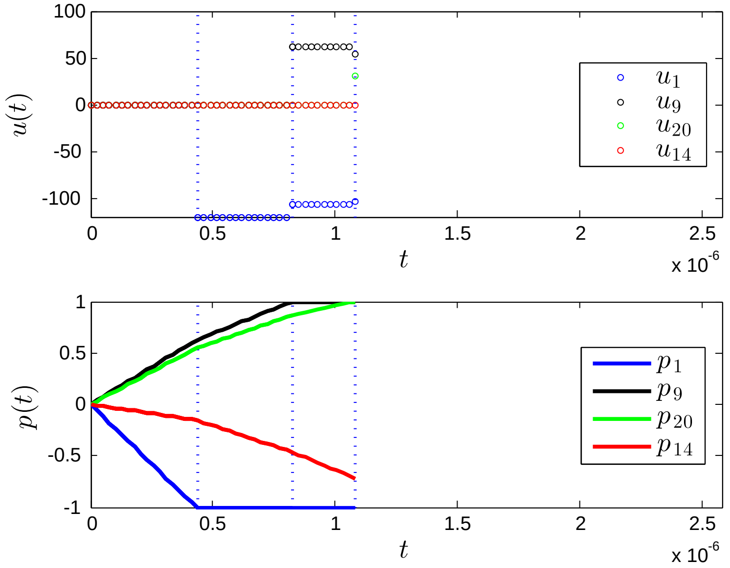
100
50
0
-50
-100
$u(t)$
0
0.5
1
1.5
2
2.5
t
x 10^{-6}
u_1
u_9
u_{20}
u_{14}
1
0.5
0
-0.5
-1
$p(t)$
0
0.5
1
1.5
2
2.5
t
x 10^{-6}
p_1
p_9
p_{20}
p_{14}

Example



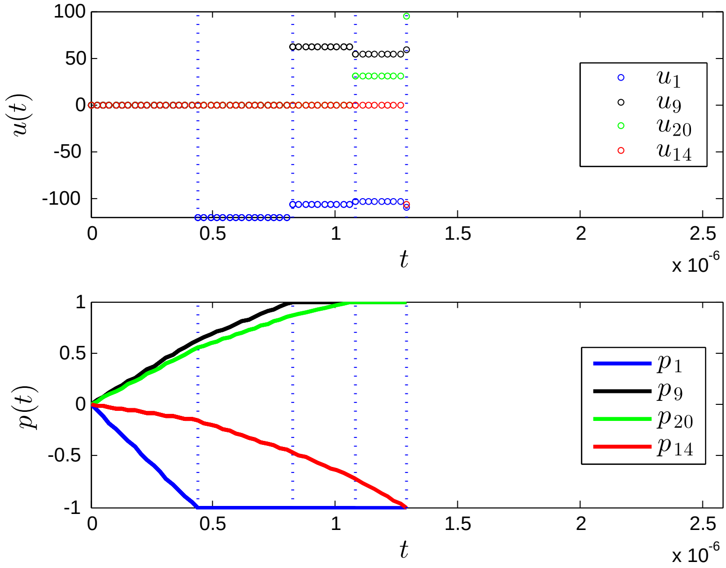
100
50
0
-50
-100
$u(t)$
0
0.5
1
1.5
2
2.5
t
x 10^{-6}
u_1
u_9
u_{20}
u_{14}
1
0.5
0
-0.5
-1
$p(t)$
0
0.5
1
1.5
2
2.5
t
x 10^{-6}
p_1
p_9
p_{20}
p_{14}

Example



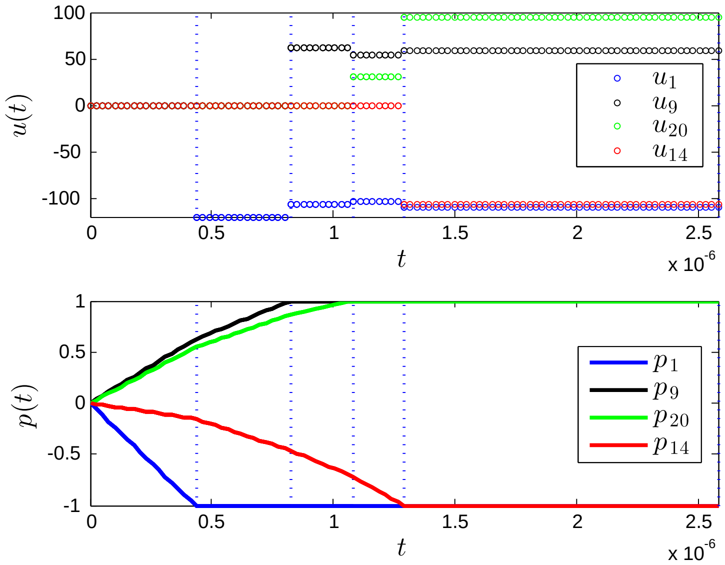
100
50
0
-50
-100
$u(t)$
0
0.5
1
1.5
2
2.5
t
x 10^-6
u_1
u_9
u_{20}
u_{14}
1
0.5
0
-0.5
-1
$p(t)$
0
0.5
1
1.5
2
2.5
t
x 10^-6
p_1
p_9
p_{20}
p_{14}

Example



100
50
0
-50
-100
$u(t)$
0
0.5
1
1.5
2
2.5
t
x 10^-6
u_1
u_9
u_{20}
u_{14}
1
0.5
0
-0.5
-1
$p(t)$
0
0.5
1
1.5
2
2.5
t
x 10^-6
p_1
p_9
p_{20}
p_{14}

Example



100
50
0
-50
-100
$u(t)$
0
0.5
1
1.5
2
2.5
t
x 10^{-6}
u_1
u_9
u_{20}
u_{14}
1
0.5
0
-0.5
-1
$p(t)$
0
0.5
1
1.5
2
2.5
t
x 10^{-6}
p_1
p_9
p_{20}
p_{14}

Relation to orthogonal matching pursuit (OMP)

Algorithm: start with $S = \emptyset$ and $u = 0$; iterate

1. add the largest entry of $A^*(b - Au)$ to S
2. set $u \leftarrow \arg\min \|b - Au\|_2^2 \quad$ s.t. $u_i = 0 \; \forall i \notin S$.

Relation to orthogonal matching pursuit (OMP)

Algorithm: start with $S = \emptyset$ and $u = 0$; iterate

1. add the <u>largest entry</u> of $A^*(b - Au)$ to S
2. set $u \leftarrow \arg\min \|b - Au\|_2^2 \quad$ s.t. $u_i = 0 \ \forall i \notin S$.

Differences:

- OMP evolves index set S;
  new method evolves ℓ_1-subgradient p, keeping more information

Relation to orthogonal matching pursuit (OMP)

Algorithm: start with $S = \emptyset$ and $u = 0$; iterate

1. add the largest entry of $A^*(b - Au)$ to S
2. set $u \leftarrow \arg\min \|b - Au\|_2^2 \quad$ s.t. $u_i = 0\ \forall i \notin S$.

Differences:

- OMP evolves index set S;
  new method evolves ℓ_1-subgradient p, keeping more information
- both add one nonzero each iteration, but new method may also drop

Relation to orthogonal matching pursuit (OMP)

Algorithm: start with $S = \emptyset$ and $u = 0$; iterate

1. add the largest entry of $A^*(b - Au)$ to S
2. set $u \leftarrow \arg\min \|b - Au\|_2^2 \quad$ s.t. $u_i = 0\ \forall i \notin S$.

Differences:

- OMP evolves index set S;
  new method evolves ℓ_1-subgradient p, keeping more information
- both add one nonzero each iteration, but new method may also drop
- both have extensions to have multiple adds/drops each iteration

Relation to orthogonal matching pursuit (OMP)

Algorithm: start with $S = \emptyset$ and $u = 0$; iterate

1. add the largest entry of $A^*(b - Au)$ to S
2. set $u \leftarrow \arg\min \|b - Au\|_2^2$ s.t. $u_i = 0 \ \forall i \notin S$.

Differences:

- OMP evolves index set S;
  new method evolves ℓ_1-subgradient p, keeping more information
- both add one nonzero each iteration, but new method may also drop
- both have extensions to have multiple adds/drops each iteration
- similar computing cost at each iteration

Numerical results: new method is more powerful than OMP at sparse recovery

Relation to LASSO

Model:

$$\min \|u\|_1 + \frac{t}{2n}\|Au - b\|_2^2$$

Optimality conditions:

$$\frac{p}{t} = A^*(b - Au), \quad p \in \partial\|u\|_1.$$

Similarities:

- $p \in \partial\|u\|_1 \cap \mathcal{R}(A^T)$, and p is continuous
- as $t \to \infty$, both u is a solution to

$$\min \|u\|_1 \quad \text{s.t. } Au = b.$$

- as t increases, both add and can also drop predictors
- sign consistency under conditions

Technical differences:

- only LASSO has an objective function

Technical differences:

- only LASSO has an objective function
- different p, except before the first predictor drop

Technical differences:

- only LASSO has an objective function
- different p, except before the first predictor drop
- different u, except at $t = 0$ and $t = \infty$

Technical differences:

- only LASSO has an objective function
- different p, except before the first predictor drop
- different u, except at $t = 0$ and $t = \infty$
- LASSO u is continuous; new u is piece-wise constant
- new method can set $u_i = 0$ *immediately*; LASSO waits for u_i to decrease 0
- LASSO+debiasing $\neq$ new method

Technical differences:

- only LASSO has an objective function
- different p, except before the first predictor drop
- different u, except at $t = 0$ and $t = \infty$
- LASSO u is continuous; new u is piece-wise constant
- new method can set $u_i = 0$ *immediately*; LASSO waits for u_i to decrease 0
- LASSO+debiasing $\neq$ new method

Qualitative differences:

- to reach the same fitting, new method requires fewer nonzeros
- given the same number of nonzeros, new method has better fitting
- LASSO is biased; new method is not

There are **differences in both variable selection and estimation**

Bias

Suppose both method select true $S = \mathrm{supp}(u^*)$.

- LASSO gives

$$\hat{u}_S(\tau) = (A_S^T A_S)^{-1} A_S b - \underbrace{\frac{m}{\tau}\mathrm{sign}(\hat{u}_S(\tau))}_{\text{bias}}$$

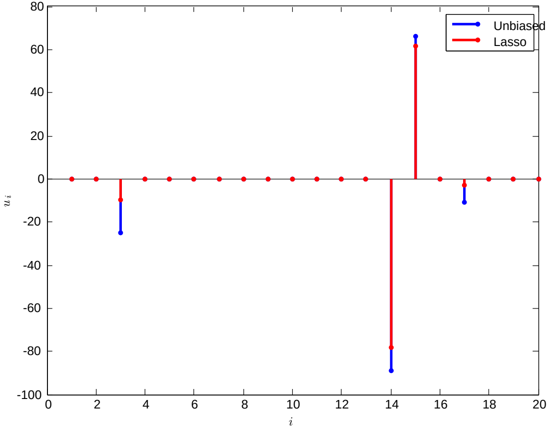
more noise $\implies$ smaller τ $\implies$ stronger bias

- new method gives

$$u_S(t) = (A_S^T A_S)^{-1} A_S b$$

- assuming 0-mean noise, $u_S(t)$ is unbiased since

$$\mathbf{E}[u_S(t)] = \mathbf{E}[(A_S^T A_S)^{-1} A_S (A_S u_S^* + \epsilon)] = u_S^*$$



Unbiased
Lasso
80
60
40
20
0
-20
-40
-60
-80
-100
0
2
4
6
8
10
12
14
16
18
20
u_i
i

Theorem

1. *For any* A *and* b, *solution to* (1) *exists;*
2. $p(t)$ *is unique and piece-wise linear;*
3. $Au(t) - b$ *is piece-wise constant;* $\|Au(t) - b\|$ *is non-increasing;*
4. *There exists a piece-wise constant* $u(t)$*;*
5. *Let* $I = \mathrm{supp}(u(t))$ *and assume 0-mean noise. Then,* $u(t)$ *is an unbiased solution to*
$$A_I u_I = b;$$
6. *There exists* t_∞ *such that for* $t \geq t_\infty$, $u(t) = u_\infty$ *is a solution to*
$$\min \|u\|_1 \quad \textit{s.t.}\ \|Au - b\|_2 = \min_w \|Aw - b\|_2.$$

Many results are essentially known from CAM 04-13 and 11-08.

Prostate tumor size

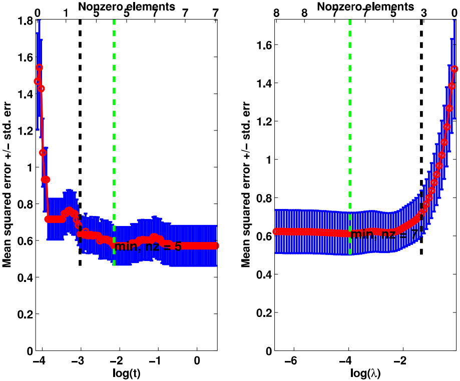
- select predictors among 8 clinical features to predict prostate tumor size
  - apply 4 different methods to 67 training cases
    - results were tested on 30 testing cases

Predictor	LS	Subset	glmnet	ISS
Intercept	2.452	2.466	2.481	2.476
lcavol	0.716	0.667	0.622	0.554
lweight	0.293	0.366	0.289	0.279
age	-0.143	0	-0.096	0
lbph	0.212	0	0.188	0.198
svi	0.310	0.268	0.262	0.238
lcp	-0.289	-0.291	-0.164	0
gleason	-0.021	0	0	0
pgg45	0.277	0.227	0.187	0.122
Test Error	0.586	0.587	0.543	0.541

LS = least squares, Subset = best subset regression

glmnet = a package with LASSO, proposed approach

Cross validation



ISS achieves better fitting with fewer nonzerso than LASSO (glmnet)

Note: exactly the same cross validation was applied to both methods

Relation to Bregman iteration

- Discretize $\dot{p} = A^*(b - Au)$ by

$$p^{k+1} = p^k + \delta A^*(b - Au^k).$$

- It is the first-order optimality condition to Bregman iteration

$$u^{k+1} \leftarrow \min D_{\|\cdot\|_1}(u; u^k) + \frac{\delta}{2n}\|Au - b\|^2,$$

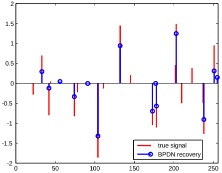
$$\text{where} \quad D_{\|\cdot\|_1}(u; u^k) := \|u\|_1 - \|u^k\|_1 - \langle p^k, u - u^k\rangle.$$

- After change of variable (CAM 04-13, 07-37)

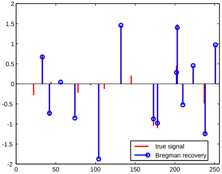
$$\begin{aligned} u^{k+1} &\leftarrow \min \|u\|_1 + \frac{\delta}{2n}\|Au - b^k\|^2, \\ b^{k+1} &\leftarrow b^k + (b - Au^k). \end{aligned}$$

- Still true if $\|\cdot\|_1$ is replace by any convex regularizer

Sparse recovery from noisy measurements



LASSO



Bregman

Relation to linearized Bregman

- Damping $\dot{p} = A^*(b - Au)$ into
$$\dot{p}(t) + \alpha \dot{u}(t) = A^*(b - Au(t)).$$
Consequence: $u(t)$ is continuous.

Relation to linearized Bregman

- Damping $\dot{p} = A^*(b - Au)$ into
$$\dot{p}(t) + \alpha\dot{u}(t) = A^*(b - Au(t)).$$
Consequence: $u(t)$ is continuous.
- Forward Euler discretization
$$p^{k+1} + \alpha u^{k+1} = p^k + \alpha u^k + \delta A^*(b - Au^k).$$

Relation to linearized Bregman

- Damping $\dot{p} = A^*(b - Au)$ into
$$\dot{p}(t) + \alpha\dot{u}(t) = A^*(b - Au(t)).$$
Consequence: $u(t)$ is continuous.
- Forward Euler discretization
$$p^{k+1} + \alpha u^{k+1} = p^k + \alpha u^k + \delta A^*(b - Au^k).$$
- Can be simplified to (in a miracle way!)
$$u^{k+1} = \alpha^{-1}\,\mathrm{shrink}(A^T y^k)$$
$$y^{k+1} = y^k + \frac{\delta}{n}(b - Au^{k+1})$$

Relation to linearized Bregman

- Damping $\dot{p} = A^*(b - Au)$ into
$$\dot{p}(t) + \alpha \dot{u}(t) = A^*(b - Au(t)).$$
Consequence: $u(t)$ is continuous.
- Forward Euler discretization
$$p^{k+1} + \alpha u^{k+1} = p^k + \alpha u^k + \delta A^*(b - Au^k).$$
- Can be simplified to (in a miracle way!)
$$u^{k+1} = \alpha^{-1} \operatorname{shrink}(A^T y^k)$$
$$y^{k+1} = y^k + \frac{\delta}{n}(b - Au^{k+1})$$
- If b is noisy, stop at a finite k for best solution.

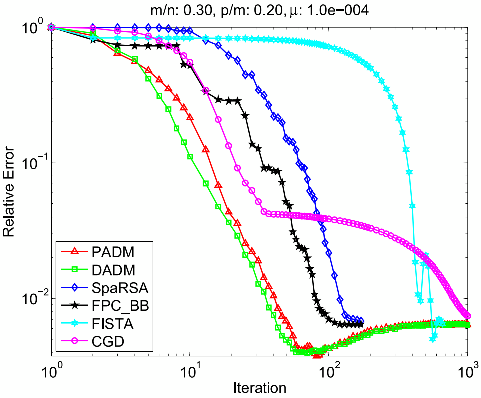
Relation to linearized Bregman

- Damping $\dot{p} = A^*(b - Au)$ into
$$\dot{p}(t) + \alpha \dot{u}(t) = A^*(b - Au(t)).$$
Consequence: $u(t)$ is continuous.
- Forward Euler discretization
$$p^{k+1} + \alpha u^{k+1} = p^k + \alpha u^k + \delta A^*(b - Au^k).$$
- Can be simplified to (in a miracle way!)
$$u^{k+1} = \alpha^{-1} \operatorname{shrink}(A^T y^k)$$
$$y^{k+1} = y^k + \frac{\delta}{n}(b - Au^{k+1})$$
- If b is noisy, stop at a finite k for best solution.
- If b is noise-free, u^k converges at a linear rate to the solution of
$$\min \|u\|_1 + \frac{\alpha}{2}\|u\|_2^2 \quad \text{s.t. } \|Au - b\|_2 = \min_w \|Aw - b\|_2.$$
Sufficiently small α (e.g., $\alpha < \frac{1}{10\|u^*\|_\infty}$ in CS) $\Longrightarrow$ u^* is an ℓ_1 minimizer

Apply different primal and dual algorithms to the same model

$$\min \|u\|_1 + \frac{t}{2n}\|Au - b\|_2^2.$$

Dual algorithms do better than the model!



Path consistency

Question: does there $\exists t$ so that solution $u(t)$ has the following properties?

- **no false positive**: if $u_i = 0$, then $u_i(t) = 0$
- **no false negative**: if $u_i \neq 0$, then $u_i(t) \neq 0$
- **sign consistency**: furthermore, $\mathrm{sign}(u) = \mathrm{sign}(u(t))$.

Path consistency

Question: does there $\exists t$ so that solution $u(t)$ has the following properties?

- **no false positive**: if $u_i = 0$, then $u_i(t) = 0$
- **no false negative**: if $u_i \neq 0$, then $u_i(t) \neq 0$
- **sign consistency**: furthermore, $\mathrm{sign}(u) = \mathrm{sign}(u(t))$.

Theorem

Under the ***Assumptions***

- *Gaussian noise:* $\omega \sim N(0, \sigma^2 I)$*,*
- *normalized column:* $\frac{1}{n} \max_j \|A_j\|^2 \leq 1$*,*

and under appropriate conditions, the new method has sign consistency and gives an unbias estimate to u^**.*

Proof is based on the next two lemmas.

No false positive

Define true support $S := \mathrm{supp}(u)$, and let $T := S^c$.

Lemma

Under ***Assumptions****, if* A_S *has full column rank and*

$$\max_{j \in T} \|A_j^T A_S (A_S^T A_S)^{-1}\|_1 \leq 1 - \eta$$

for some $\eta \in (0, 1)$*, then with high probability*

$$\mathrm{supp}(u(s)) \subseteq S, \quad \forall s \leq \bar{t} := O\left(\frac{\eta}{\sigma}\sqrt{\frac{m}{\log n}}\right).$$

Proof uses: (i) concentration inequality and (ii) if $\mathrm{supp}(u(s)) \subseteq S, \; s \leq t$, then

$$p(s)_T = A_T^T A_S (A_S^T A_S)^{-1} p(s)_S + t A_T^* P_{A_S^\perp} w, \quad s \leq t.$$

No false negative / sign consistency

Lemma

*Under **Assumptions**, if $A_S^* A_S \succeq \gamma I$ and*

$$u_{\min} \geq \max\left\{ O\left(\frac{\sigma}{\sqrt{\gamma}} \sqrt{\frac{\log |S|}{m}} \right), O\left(\frac{\sigma \log |S|}{\eta\gamma} \sqrt{\frac{\log n}{m}} \right) \right\},$$

then there exist t^* (which can be given explicitly) so that with high probability

$$\mathrm{sign}(u(t)) = \mathrm{sign}(u)$$

and $u(t) = u_S - (A_S^* A_S)^{-1} A_S^* \omega$ obeys

$$\|u(t) - u\|_\infty \leq u_{\min}/2.$$

- first term in $\max$ ensures $\|(A_S^* A_S)^{-1} A_S^* \omega\|_\infty \leq u_{\min}/2$
- second term ensures: $\inf\{t : \mathrm{sign}(u_S(t)) = \mathrm{sign}(u_S)\} \leq \bar{t}$.