15.095: Machine Learning under a Modern Optimization Lens

Lecture 2: Sparse Linear Regression

Outline

Best Subset Selection

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbf{X} \in \mathbb{R}^{n \times p}$$

Least squares:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

(ℓ_2-)regularized least squares ("ridge regression"):

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2$$

Best Subset Selection:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\ \text{s.t.} \quad & \|\boldsymbol{\beta}\|_0 \leq k, \end{aligned}$$

where $\|\boldsymbol{\beta}\|_0 = \sum_i \mathbf{1}_{\beta_i \neq 0}$ = number of nonzeroes of $\boldsymbol{\beta}$.

ℓ_p norm ($p \in [1, \infty]$): $\|\boldsymbol{\beta}\|_p = \left(\sum_i |\beta_i|^p\right)^{1/p}$

Best Subset Selection

$$\min_{\boldsymbol{\beta}} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \ \text{ subject to } \ \|\boldsymbol{\beta}\|_0 \leq k$$

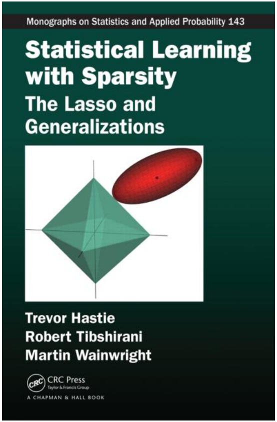
How solved historically?

- Enumeration: Furnival and Wilson (1974) solve it by implicit enumeration, `leaps` routine in `R`. Cannot scale beyond $p = 30$.
- Convex relaxation: Lasso was proposed in 1996 by Tibshirani (24,000+ citations):

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_i |\beta_i|.$$

- Candès–Tao: Under regularity conditions on $\mathbf{X}$, Lasso leads to sparse models and good predictive performance.

Lasso



Widely held belief: statistical problems with discrete elements are intractable, and convex optimization is our only hope.

MIO Approach

Key idea: Use *Mixed Integer Optimization* (MIO) modeling techniques to capture discrete nature of optimization problem

Introduce auxiliary variables $\mathbf{z}$ with

$$z_i \in \{0, 1\}, \quad \beta_i \neq 0 \implies z_i = 1$$

Can be expressed in several different ways:

1. "Big M" constraint:
$$|\beta_i| \leq M_i \cdot z_i$$
2. Special ordered set constraint (SOS-1):
$$(1 - z_i)\beta_i = 0$$

MIO Approach

Key idea: Use *Mixed Integer Optimization* (MIO) modeling techniques to capture discrete nature of optimization problem

$$
\begin{aligned}
\min_{\boldsymbol{\beta},\mathbf{z}} \quad & \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \\
\text{s.t.} \quad & |\beta_i| \leq M_i \cdot z_i, i = 1, \ldots, p \\
& \sum_{i=1}^{p} z_i \leq k \\
& z_i \in \{0, 1\}, i = 1, \ldots, p.
\end{aligned}
$$

MIO Approach—Setting M

Only thing left is to set values of M_i.

For the case $n > p$,

$$\begin{aligned} u_i^+ :=\ & \max_{\boldsymbol{\beta}} \ \beta_i \\ \text{s.t.}\ & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \leq \mathrm{UB} \end{aligned} \quad \text{and} \quad \begin{aligned} u_i^- :=\ & \min_{\boldsymbol{\beta}} \ \beta_i \\ \text{s.t.}\ & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \leq \mathrm{UB}, \end{aligned}$$

where UB is an upper bound to the best subset problem problem.

$M_i = \max\{|u_i^+|, |u_i^-|\}$ serves as an upper bound to $|\widehat{\beta}_i|$.

Data Set

Diabetes data set—442 patients with ten baseline measurements:

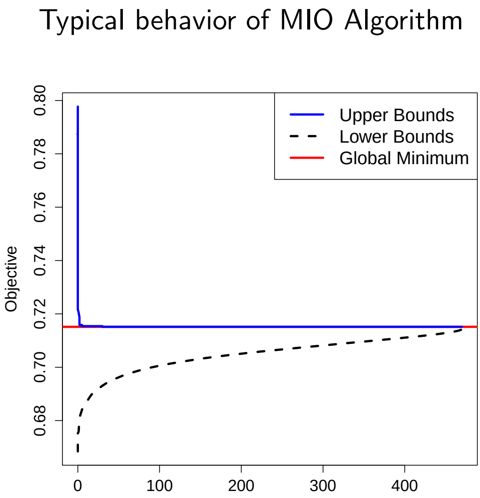
- age, sex, body mass index (BMI), average blood pressure
- six blood serum measurements

Interested in predicting hemoglobin measure in one year's time

Random subsample of $n = 350$ patients with $p = 64$ variables:

- 10 original variables
- $\binom{10}{2} = 55$ second-order interaction variables of form $x_i \cdot x_j$
- Variable x_{sex}^2 removed (because $x_{\text{sex}}^2 = x_{\text{sex}}$)

Typical MIO Behavior



Diabetes Dataset ($n = 350, p = 64, k = 6$)

Overall Strategy for solving MIO

Warm starts via first order methods—finding good feasible solutions

Improved formulations

First Order Method

Consider

$$\min_{\boldsymbol{\beta}} \quad g(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \ \text{ subject to } \ \|\boldsymbol{\beta}\|_0 \leq k$$

Note that g is convex and

$$\|\nabla g(\boldsymbol{\beta}) - \nabla g(\boldsymbol{\beta}_0)\|_2 \leq \ell \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2.$$

("g has ℓ-Lipschitz gradients.")

This implies that for any $L \geq \ell$,

$$g(\boldsymbol{\beta}) \leq Q(\boldsymbol{\beta}) = g(\boldsymbol{\beta}_0) + \langle \nabla g(\boldsymbol{\beta}_0), \boldsymbol{\beta} - \boldsymbol{\beta}_0 \rangle + \frac{L}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2.$$

For the purpose of finding feasible solutions, we propose

$$\min_{\boldsymbol{\beta}} \quad Q(\boldsymbol{\beta}) \ \text{ s.t. } \ \|\boldsymbol{\beta}\|_0 \leq k.$$

Solution

How does this help us? This upper bound can be solved in closed form!

$$\min_{\|\boldsymbol{\beta}\|_0 \leq k} Q(\boldsymbol{\beta}) \quad \equiv \quad \min_{\|\boldsymbol{\beta}\|_0 \leq k} \ \frac{L}{2}\|\boldsymbol{\beta} - \underbrace{(\boldsymbol{\beta}_0 - \nabla g(\boldsymbol{\beta}_0)/L)}_{=:\mathbf{u}}\|_2^2 - \frac{1}{2L}\|\nabla g(\boldsymbol{\beta}_0)\|_2^2.$$

Reduces to

$$\min_{\|\boldsymbol{\beta}\|_0 \leq k} \|\boldsymbol{\beta} - \mathbf{u}\|_2^2.$$

Optimal solution is $\boldsymbol{\beta}^* = H_k(\mathbf{u})$, where $H_k(\mathbf{u})$ retains the k largest magnitude elements of $\mathbf{u}$ and sets the rest to zero.

First Order Algorithm

Algorithm 1

Input: $g(\boldsymbol{\beta})$, L, ϵ.

Output: A first order stationary solution $\boldsymbol{\beta}^*$.

1. Initialize with $\boldsymbol{\beta}_1 \in \mathbb{R}^p$ such that $\|\boldsymbol{\beta}_1\|_0 \leq k$.
2. For $m \geq 1$
$$\boldsymbol{\beta}_{m+1} \in H_k \left(\boldsymbol{\beta}_m - \frac{1}{L} \nabla g(\boldsymbol{\beta}_m) \right)$$
3. Repeat Step 2, until $g(\boldsymbol{\beta}_m) - g(\boldsymbol{\beta}_{m+1}) \leq \epsilon$.

Rate of Convergence

The sequence $g(\boldsymbol{\beta}_m)$ converges to $g(\overline{\boldsymbol{\beta}})$ where

$$\overline{\boldsymbol{\beta}} = H_k\left(\overline{\boldsymbol{\beta}} - \frac{1}{L}\nabla g(\overline{\boldsymbol{\beta}})\right).$$

After M iterations:

$$\min_{m=0,\ldots,M} \|\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m\|_2^2 \leq \frac{2\cdot\left(g(\boldsymbol{\beta}_0) - g(\overline{\boldsymbol{\beta}})\right)}{M\cdot(L-\ell)}$$

After $M = O(1/\epsilon)$ iterations Algorithm 1 converges.

Quality of Solutions

Diabetes data: $n = 350,\ p = 64$.

Relative Accuracy $= (f_{\text{alg}} - f_*)/f_*$

maximum time of 500 seconds

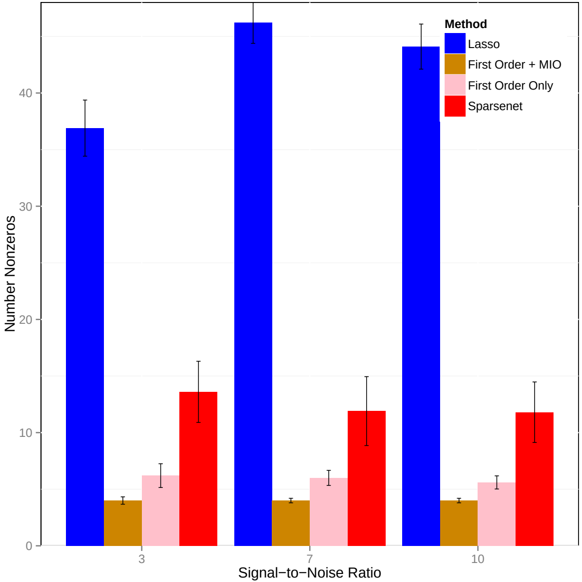
k	First Order Accuracy	First Order Time	MIO Cold Start Accuracy	MIO Cold Start Time	MIO Warm Start Accuracy	MIO Warm Start Time
9	0.1306	1	0.0036	500	0	346
20	0.1541	1	0.0042	500	0	77
49	0.1915	1	0.0015	500	0	87
57	0.1933	1	0	500	0	1

Computational experiments

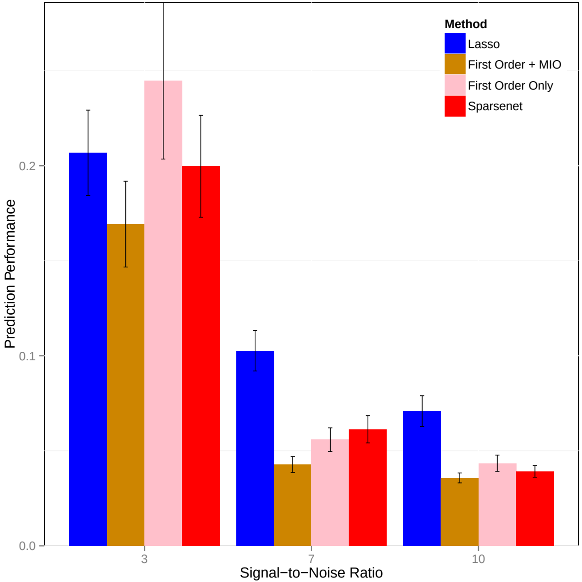
Comparison of various methods:

- Lasso
- Algorithm 1 + MIO
- Algorithm 1 on its own
- Sparsenet (another popular method)

Sparsity Detection for $n = 50$, $p = 2000$



Prediction Error for $n = 50$, $p = 2000$



Improving formulation strength

Additional improvements to the MIO model

$$v_i^+ := \max_{\boldsymbol{\beta}} \ \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \qquad\qquad v_i^- := \min_{\boldsymbol{\beta}} \ \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle$$
$$s.t. \quad \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 \leq \mathrm{UB}. \qquad\qquad s.t. \quad \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 \leq \mathrm{UB}.$$

$v_i = \max\{|v_i^+|, |v_i^-|\}$ serves as an upper bound to $|\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle|$

Add constraints $\|\mathbf{X}\widehat{\beta}\|_\infty \leq \max_i v_i$ and $\|\mathbf{X}\widehat{\beta}\|_1 \leq \sum_i v_i$ to the model.
$\hookrightarrow$ Does not change solutions, but improves formulation strength.

A Dual Perspective

Consider Best Subset Selection with ridge objective:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \\ \text{s.t.} \quad & \|\boldsymbol{\beta}\|_0 \leq k. \end{aligned}$$

Letting $S := \{\mathbf{s} \in \{0,1\}^p \ : \ \mathbf{1}'\mathbf{s} \leq k\}$, we can rewrite this as

$$\min_{\mathbf{s} \in S} \left[\min_{\boldsymbol{\beta}_s \in \mathbb{R}^k} \|\mathbf{y} - \mathbf{X}_s \boldsymbol{\beta}_s\|_2^2 + \lambda \|\boldsymbol{\beta}_s\|_2^2 \right].$$

Solution:

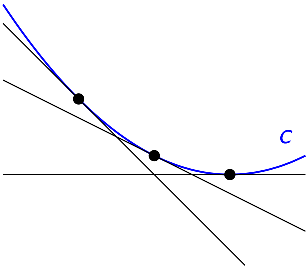
$$\begin{aligned} \min \quad & c(\mathbf{s}) = \mathbf{y}' \left(\mathbf{I}_n + \frac{1}{\lambda} \sum_j s_j \mathbf{K}_j \right)^{-1} \mathbf{y} \\ \text{s.t.} \quad & \mathbf{s} \in S, \end{aligned}$$

where $\mathbf{K}_j := \mathbf{X}_j \mathbf{X}_j'$.
$\hookrightarrow$ Binary convex optimization problem!

Using Convexity

By convexity of c, for any $\mathbf{s}, \bar{\mathbf{s}} \in S$,

$$c(\mathbf{s}) \geq c(\bar{\mathbf{s}}) + \sum_i \frac{\partial c(\bar{\mathbf{s}})}{\partial s_i} \cdot (s_i - \bar{s}_i)$$



A Cutting Plane Algorithm

This leads to a cutting plane algorithm:

1. Pick some $\mathbf{s}_1 \in S$ and set $C_1 = \{\mathbf{s}_1\}$.

2. For $t \geq 1$, solve

$$z_t^* \quad = \quad \min_{\mathbf{s} \in S} \left[\max_{\bar{\mathbf{s}} \in C_t} c(\bar{\mathbf{s}}) + \sum_i \frac{\partial c(\bar{\mathbf{s}})}{\partial s_i} \cdot (s_i - \bar{s}_i) \right].$$

3. If solution $\mathbf{s}_t^*$ to Step 2 has $c(\mathbf{s}_t^*) > z_t^*$, then set $C_{t+1} := C_t \cup \{\mathbf{s}_t^*\}$ and go back to Step 2.

Scalability

Cutting plane algorithm can be faster than Lasso.

		Exact T [s]			Lasso T [s]		
		$n = 10\text{k}$	$n = 20\text{k}$	$n = 100\text{k}$	$n = 10\text{k}$	$n = 20\text{k}$	$n = 100\text{k}$
$k = 10$	$p = 50\text{k}$	21.2	34.4	310.4	69.5	140.1	431.3
	$p = 100\text{k}$	33.4	66.0	528.7	146.0	322.7	884.5
	$p = 200\text{k}$	61.5	114.9	NA	279.7	566.9	NA
$k = 20$	$p = 50\text{k}$	15.6	38.3	311.7	107.1	142.2	467.5
	$p = 100\text{k}$	29.2	62.7	525.0	216.7	332.5	988.0
	$p = 200\text{k}$	55.3	130.6	NA	353.3	649.8	NA
$k = 30$	$p = 50\text{k}$	31.4	52.0	306.4	99.4	220.2	475.5
	$p = 100\text{k}$	49.7	101.0	491.2	318.4	420.9	911.1
	$p = 200\text{k}$	81.4	185.2	NA	480.3	884.0	NA

Phase Transitions

- $\mathbf{Y} = \mathbf{X}\beta_{\text{true}} + \mathbf{E}$ where $\mathbf{E}$ is zero mean noise uncorrelated with the signal $\mathbf{X}\beta_{\text{true}}$.

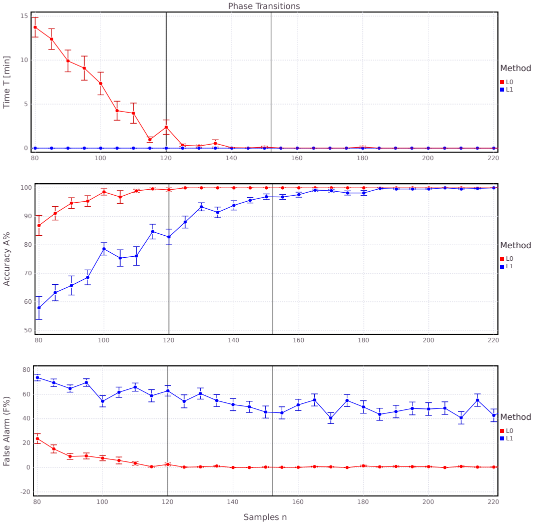
- Accuracy and false alarm rate of a certain solution $\beta^\star$

$$A\% := 100 \times \frac{|\operatorname{supp}(\beta_{\text{true}}) \cap \operatorname{supp}(\beta^\star)|}{k}$$

$$F\% := 100 \times \frac{|\operatorname{supp}(\beta^\star) \setminus \operatorname{supp}(\beta_{\text{true}})|}{|\operatorname{supp}(\beta^\star)|}.$$

- Perfect support recovery occurs only then when $\beta^\star$ tells the whole truth ($A\% = 100$) and nothing but the truth ($F\% = 0$).

Phase Transitions



Remark on Complexity

- Traditional complexity theory suggests that the difficulty of a problem increases with dimension.
- Sparse regression problem has the property that for small number of samples n, the dual approach takes a large amount of time to solve the problem, but most importantly the optimal solution does not recover the true signal.
- However, for a large number of samples n, dual approach solves the problem extremely fast and recovers 100% of the support of the true regressor $\boldsymbol{\beta}_{\text{true}}$.

Summary

The widely held belief that statistical problems of a discrete nature are intractable needs revision.

Advances in modern MIO techniques allow us to solve large scale instances, in some settings *even faster* than using convex techniques alone.

An example of general methodological approach for the class: reexamine old statistical problems and bring a new perspective by using all of the current knowledge in optimization.

References

- "Best subset selection via a modern optimization lens," Bertsimas, King, and Mazumder, *Annals of Statistics*.
- Dual perspective was first used in "Sparse learning via Boolean relaxations," Pilanci, Wainwright, and El Ghaoui, *Mathematical Programming, Series B*. Computational results can be found in Bertsimas and van Parys (2017).